Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575 *Bahar* 2023 *Spring* 2023 *Cilt: 14-Sayı: 1* *Volume: 14-Issue: 1*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

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

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

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

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

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

Hakem Kurulu / Referee Board

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

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Mart 2023, Sayı: 14-1)

Journal of Measurement and Evaluation in Education and Psychology (March 2023, Issue: 14-1)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 1-18

Differential Bundle Functioning of National Examinations Council Mathematics Test Items: An Exploratory Structural Equation Modelling Approach

Oluwaseyi Aina OPESEMOWO* Musa Adekunle AYANWALE**
Titilope Racheal OPESEMOWO*** Eyitayo Rufus Ifedayo AFOLABİ****

Abstract

A differential bundle function (DBF) is a situation in which examinees who are of the same ability but are from different groups are required to answer groups of items differently. DBF undermines the validity of the instrument if inadequately considered. The study examines the dimensionality of the 2017 NECO Mathematics items, determines the effect of DBF on 2017 Mathematics items concerning sex, and investigates the effect of DBF on 2017 Mathematics items concerning school ownership. This study explores Exploratory Structural Equation Modelling (ESEM), which permits the cross-loading of items that are not allowed in other models. The ex-post facto research design was adopted using secondary data, while six bundles were generated via the specification table. The population for the study comprised all 1,034,629 Senior School three students. A total of 815,104 students were selected using the simple random technique. The instrument for the study was 2017 NECO Mathematics paper three with a Cronbach's alpha of 0.82, and data were analysed using Mplus 7.4. The results revealed that the 2017 NECO Mathematics is multidimensional and items in the bundles possess construct validity as they functioned differentially to examinees' sex and school type. We recommend ESEM has a better approach to examining DBF on 2017 NECO Mathematics test items.

Keywords: Differential bundle functioning, exploratory structural equation modelling, National Examinations Council

Introduction

The dismal performance of examinees in the Senior School Certificate Examination (SSCE) could be a result of differential bundle performance that is spotted among examinees' group (male/female), and this (dismal performance) can lead to item bundle bias. To ensure that item bundles are fair to all intended groups, examination bodies should modify or delete bundles that may flag Differential Bundle Functioning (DBF) across examinees' groups. The instrument to measure the ability of groups of examinees becomes unfair when DBF occurs. There are four ways by which test fairness is categorised as submitted by standards for Educational and Psychological Testing (Boughton et al., 2000). Firstly, a fair test must be free from bias. Bias occurs when tests yield or promote scores that result in different meanings for members of different groups of examinees with the same competence level. Secondly, test fairness requires that examinees have received equal justice and treatment in the testing process. The achievement of fair treatment in a standardised test can be actualised when awarding scores to individuals and examinees groups by considering the items in the test and the testing context. Third of them is that test outcomes must be equitable to ensure test fairness, meaning examinees must have an equal opportunity to demonstrate proficiency in the measured construct. Examinees with the same

* Postdoctoral Research Fellow, University of Johannesburg, Faculty of Education, Johannesburg-South Africa, opesemowoo@gmail.com, ORCID ID: 0000-0003-0242-7027
** Senior Research Fellow, University of Johannesburg, Faculty of Education, Johannesburg-South Africa, ayanwalea@uj.ac.za, ORCID ID: 0000-0001-7640-9898
*** Research Assistant, Obafemi Awolowo University, Faculty of Education, Ile Ife-Nigeria, oluwatimilehint@gmail.com, ORCID ID: 0000-0002-0553-7355
**** Prof., Obafemi Awolowo University, Faculty of Education, Ile Ife-Nigeria, eriafolabi@gmail.com, ORCID ID: 0000-0002-0014-0711

To cite this article:
Opesemowo, O. A., Ayanwale, M. A., Opesemowo, T. R., & Afolabi, E. R. I. (2023). Differential bundle functioning of National Examinations Council mathematics test items: An exploratory structural equation modelling approach. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 1-18. https://doi.org/10.21031/epod.1142713

Received: 9.7.2022
Accepted: 8.11.2022

capacity should receive the same score if there is no bias in the testing process. Lastly, to be fair to examinees, it is important to ensure they have had the opportunity to learn the content covered in the achievement domain during the preparation for the exam (Boughton et al., 2000). Summarily, the multifaceted nature of test fairness has made it practically impossible to have a generally accepted definition. When groups or bundles of items are unfair in measuring the same construct, such groups or bundles of items reflect DBF.

DBF occurs after controlling for the overall capacity of the construct measured by the test for examinees with the same ability but belonging to different groups who have a different probability of answering groups of items correctly. Item bias and impact may be responsible for DBF (Latifi et al., 2016). More so, when it involves two groups, such as examinees from rural/urban communities, male/female examinees, or public/private school students with the same ability, one would expect that examinees receive a similar score on a particular bundle of items. The comparability of test results across cultures has also been investigated using DBF (Ong et al., 2015). When one group persistently receives a lower score on an item bundle because of insufficient knowledge to answer items correctly in the bundle or something other than the knowledge of the subject matter being measured, DBF is said to have taken place.

Similarly, examinees of the same ability in this comparison group (male/female, urban/rural) are expected to answer the clustered items correctly and receive the same score for the correctly answered item bundle. But when the contrary occurs, bias is said to have been introduced against a particular group. Furthermore, for the bundle of items to measure mathematics ability correctly, individuals who have similar knowledge and expertise should have an equal chance of getting the answer correctly. When people with the same capacity in different groups, say male and female have a different probability of successfully answering an item, that item is said to function differently (differential performance) across groups (Ong et al., 2015). Differential performance may be assessed for an individual item called Differential Item Functioning (DIF). However, DIF occurs when an examinee with the same ability but belongs to different groups has a different probability of answering an item correctly.

In contrast, when it involves groups of items measuring the same construct and examinees with the same ability have a different chance of responding correctly to such bundles, DBF occurs. When the probability or chance to answer bundles of items rightly differs from examinees with the same ability level but belong to a separate group, DBF takes place (Min & He, 2020). The concept of DBF was built upon the DIF. In the DBF, items will be categorised into bundles and crisscross whether any item in the bundle demonstrates differential performance. To bundle items, various researchers (Furlow et al., 2009; Gierl et al., 2005; Oshima et al., 1998) have outlined diverse methods to group items into bundles.

Item Bundle Creation

Item bundle is a set of test items that are supposed to measure a universal secondary dimension e.g., items measuring the same construct. In addition, a bundle is a set of items measuring the same construct or the measurement of groups of items to test a particular learning domain (cognitive, affective, and psychomotor). An established principle guides item bundle creation. The DBF impacts ability estimation in no small measure, which is why bundling items suspected of DBF is crucial. If such DBF is not detected, there can be bias in both measurements and ability estimations. A bundle can be created using several organizing principles. These include a table of specifications, expert knowledge, and statistical detection. Based on the test's different content and cognitive dimensions, the table of specifications indicates a multidimensional structure in the data. Thus, items can be sampled from this specification table to determine whether different content areas have multiple dimensions. Expert knowledge is a confirmatory approach (Douglas et al., 1996). To use this method, content experts will be required to identify groups of items that are believed to measure the same construct in the test. With this method, the expert will examine each item and then determine if the items share a common theme or similar content to make bundles to test for DBF based on these themes or similarities in content. The exploratory approach to bundling items has also been proposed by (Douglas et al., 1996). This approach involves

using statistical procedures to identify distinct dimensions; however, various analytical methods are available for structuring items as a group, such as factor analysis, cluster analysis, and multidimensional scaling, to name just a few.

After reviewing the methods of creating bundles, the researchers adopted a table of specifications in this study. Likewise, various statistical methods such as the Simultaneous Item Bias Test (Shealy & Stout, 1993; Walker et al., 2011) and Multiple Indicator Multiple Causes (Finch, 2012; Lee et al., 2016; Montoya & Jeon, 2019; Mucherah et al., 2012) have been used to investigate DBF. The Exploratory Structural Equation Modeling (Asparouhov & Muthen, 2009) was adopted in this study. In Africa and Nigeria, various research studies have been conducted on DIF, but there has been a dearth of research on DBF. Consequently, DBF poses significant threats to item and bundle parameters that inform sex and other examinees' characteristics on NECO Mathematics test item performance. Such risks reflect noticeably on examinees' performance in Mathematics and may be responsible for the dismal performance commonly reported; therefore, the need arises to illuminate this threat of DBF using the Exploratory Structural Equation Modeling (ESEM) approach. It is imperative to state that this study employs the ESEM to determine if DBF exists among examinees' sex and school type in NECO Mathematics items. This examination is peculiar because all students who aspire to proceed to the higher institution of learning in Nigeria must pass the certification examination.

Appraisal of ESEM

Before the introduction of ESEM (Asparouhov & Muthén, 2009; Marsh et al., 2014; Morin & Maïano, 2011), the Exploratory Factor Analysis (Jennrich & Sampson, 1966), and Confirmatory Factor Analysis (Jöreskog, 1969) have been used to test factor, convergent, and divergent scores of numerous psychological instruments. Exploratory Factor Analysis (EFA) seeks to uncover the underlying structure of a relatively large set of variables. This was the first technique that was commonly used for factor analysis. The EFA is applied at an early stage of instrument development, and at this stage, there is the nonappearance of the structural specification of the instrument (Tsigilis et al., 2018). The researcher described this as a data-driven technique (Brown, 2015). The EFA has some limitations. One of these limitations is that it does not include technique effect adjustments. For instance, two items with similar wording can appear in an instrument. Therefore, we often need to include residual correlations to explain the covariance of these items with their latent constructs. Researchers often take into account the comparison between scores obtained from different groups of participants when analysing instrument scores, as well. It can be concluded, therefore, that the comparison scores have meaning only if the same number is interpreted the same way for all the groups in a particular study.

On the other hand, according to pre-established theory, a Confirmatory Factor Analysis (CFA) attempts to determine whether the number of factors and the loading of measured variables corresponds to what was expected based on the number of factors and loadings. The CFA is based on the assumption that items are loaded onto their respective factors, and those cross-loading items onto one or more latent factors are not permitted. An instrument structure in CFA can be determined by looking at the theoretical assumptions as well as the outcomes of previous EFAs or results that were produced. CFA has therefore been described as a methodological approach that is conceptually driven (Tsigilis et al., 2018). It is worth noting that the advantage of CFA is its ability to elucidate whether and to what extent the measurement model generalises across groups, as well as the relative consistency of the scores obtained. Several modifications have been introduced into the exploratory model to improve the model's fit by examining certain aspects of the model that are ill-fit to the data. The ESEM method was developed by Asparouhov and Muthén (2009) as an improvement to the previous approach. The ESEM allows the user to combine both the EFA and CFA in one model, providing a holistic framework that allows both to be used simultaneously. There is no doubt that ESEM is an improvement on EFA and CFA in that it combines both improvements into a single framework where factors will cross-load at some point. ESEM also has the advantage of allowing the simultaneous analysis of all cross-loadings in the form of a single cross-loading at a time, which can be calculated based on the modifications indices of the cases that have been analysed (Morin & Maïano, 2011) in a single step. More importantly, when compared with EFA and CFA, ESEM is much more accurate at fitting the data to the model when compared to

both. So far, several studies have been conducted on DBF using other models in other countries. Still, there is a paucity of research on DBF using ESEM in Nigeria to the best of our knowledge.

Purpose of the Study

Examinees continue to perform dismally in NECO mathematics, and major stakeholders in the education industry continue to pay attention to the issue. It has been attempted by several researchers to identify the factors responsible (such as shortage of qualified teachers (Ojimba, 2012), lack of equipment and instructional materials for effective teaching (Akale, 1997), poorly motivated teachers, and overcrowded classrooms (Asikhia, 2010), students' poor attitude towards mathematics (Akinsola, 1994), poor methods of teaching mathematics (Asikhia, 2010), poor learning environment (Black, 2001; Tata et al., 2014), students poor study habits and orientation (Aremu & Sokan, 2003; Umameh, 2011), school location (Adeyemo, 2005), lack of parental participation (Uwadie, 2012), gender of the teacher (Adeyegbe & Oke, 2002; Adeyemo, 2005), nature of the test items and examinees' characteristics (Ayanwale, 2019; Awopeju & Afolabi, 2016; Adeyemo & Opesemowo, 2020) for this performance of examinees in the external exam. There have also been several studies that suggest ways to improve student's math performance, such as mathematics can be taught in indigenous languages (Adegoke, 2011), improving instructional techniques (Abina, 2014), remunerating teachers well, and creating a conducive learning environment (Uwadie, 2012). Even though researchers have provided several interventions to improve performance, this dismal trend still persists. As a contributing factor, DBF was investigated in this study, which is different from what has been studied by others. There is a need to address bias since tests are used as gatekeepers for educational opportunities, and test items should be fair to all students. A test is relevant only if it produces valid outcomes for different subpopulations with the same measures. In addition, the importance of ensuring fairness and equity among examinees cannot be overstated. It is important to provide equal opportunities for all examinees to display their knowledge and perform well according to their demographic profiles (Ayanwale, 2022). When developing their test items, does NECO take this situation into consideration? To the best of the researcher's knowledge, this remains a mystery. When a test contains DBF elements, the student's performance will be adversely affected. Therefore, it is essential to examine the DBF of this public examining body from various demographic perspectives. Specifically, the purpose of this study was to ascertain the dimensionality of the 2017 NECO mathematics items, as well as determine the impact of DBF on 2017 NECO mathematics items based on school ownership and gender.

Research Questions

Research questions addressed in this study include:

1. What is the dimensionality of the 2017 NECO Mathematics items?
2. What is the effect of DBF on 2017 NECO Mathematics items concerning sex?
3. Is there any influence of DBF on 2017 NECO Mathematics items concerning school ownership?

Method

Design

This is quantitative research using ex-post facto design. As a design, ex post facto is known as "after-the-fact" research and examines how an independent variable (groups with certain qualities that already exist before a study) influences a dependent variable. As a result, a researcher cannot modify or manipulate actions or behaviours that have already taken place or specific traits and characteristics that a participant has (Creswell, 2003). Data were drawn from candidates' responses who wrote the 2017 NECO Mathematics paper three examinations. Mathematics paper 3 consists of multiple-choice items. The population for the study comprised all of the 1,034,629 Senior School three (SS 3) students who

registered and took the examination. The population figure was made available in the data provided by NECO. The NECO Mathematics examination is a national examination usually administered annually and taken by all candidates in the 36 states, including the Federal Capital Territory (FCT), Nigeria.

Participants

A survey system sample calculator was used to determine the sample size. The sample size was set at a 95% confidence level and 0.05 confidence interval. The study sample consisted of 815,104 SS 3 students, 393,695 (48.3%) males, and 421,409 (51.7%) females were selected using a simple random technique. Also, 497 schools (i.e., 318 (63.98%) private and 179 (36.02%) public schools) across the six geopolitical zones in Nigeria that enrolled students for the NECO Mathematics examination were selected using purposive sampling techniques. Data were retrieved from the Optical Mark Recorder (OMR) sheets, obtained from the NECO head office, Minna, Niger State, Nigeria.

Instrument

The 2017 June/July NECO SSCE Mathematics paper three examination was the instrument. It was a dichotomous (i.e., correct response scored one while wrong response scored zero) multiple-choice examination comprising 60 items with a key and four distracters making five alternative responses, and the items were based on the Senior Secondary School (SSS) Mathematics curriculum. Examinees had to provide information about themselves, such as sex, location, name, examination number, school name, serial number, and subject code. The response options for the instrument range from letters A-E. After the third year of SSS, the SSCE is usually administered. In much the same way, the NECO exam serves as an assessment mechanism that ascertains the extent to which a student has acquired essential skills and competencies. A specifications table (Table 1) showed how items were distributed across the behavioural objectives and contents. The instrument has a Cronbach alpha of 0.89 reliability.

The table of specifications (Table 1) demonstrated that knowledge possesses seven items representing 11.6%, comprehension had six items with 10%, the analysis had 22 items representing 36.7%, the application had 16 items cum 26.7%, synthesis had five items representing 8.3%, and evaluation had four items representing 6.7%. Also, it may deduce that analysis revealed the highest number of items while evaluation had the least items. Similarly, number and numeration showed 11 items representing 18.3%, algebra had 18 items accounting for 30%, mensuration showed 6 items representing 10%, geometry had 9 items representing 15%, statistics and probability had 10 items showing 16.7% and introduction to calculus had 6 items which accounted for 10%. Additionally, it was shown that algebra had the highest number of items, while mensuration and introduction to calculus possessed the least number of items.

Data Analysis

Data obtained was analysed using Mplus software version 7.4 (Muthén & Muthén, 2012) and estimated with the robust maximum likelihood estimator (MLR), which provides standard errors and tests of model fit that are robust to the non-normality of the data. Also, examinees' responses were subjected to the Stout test of essential unidimensionality (Stout, 1987), a nonparametric analysis using DIMTEST package.

Table 1

Table of Specification of the 2017 NECO Mathematics Items

Content	**Cognitive Skills**						**Total**
	Know. (11.6%)	Comp. (10%)	Ana. (36.7%)	App. (26.7%)	Syn. (8.3%)	Eva. (6.7%)	
N/N (18.3%)	0	2(ITs 20 & 33)	1(IT 9)	2(ITs 5 & 10)	4(ITs 1,3,4&11)	2(ITs 2&6)	11
ALG. (30%)	1(IT 24)	3(ITs 18,31&48)	6(ITs 8,15,19,23,30&45)	8(ITs 7,12,13,16,21,22,25&32)	0	0	18
MEN. (10%)	0	0	4(ITs 37,38,39, &42)	2(ITs 17&43)	0	0	6
GEO. (15%)	0	0	8(ITs 34,35,36,40,41,44,46&51)	1(IT 48)	0	0	9
STAT/PROB. (16.7%)	6(Its 26,27,50,52&53)	1(ITs 49)	3(ITs 54,55&56)	0	0	0	10
INTRO. TO CAL. (10%)	0	0	0	3(ITs 14,28&29)	1(IT 59)	2(ITs 58&60)	6
TOTAL	7	6	22	16	5	4	60

Note. IT = Item; ITs = Items; N/N = Number and Numeration; ALG = Algebra; MEA= Mensuration; GEO = Geometry; STAT/PROB = Statistics and Probability; INTO. TO CAL. = Introduction to Calculus; Know.: Knowledge; Comp. = Comprehension; Ana. = Analysis; App. = Application; Syn. = Synthesis; Eva. = Evaluation

Model Fit Statistics

Several model fit statistics have been used by researchers to assess structural equation models, but in this study, the researchers considered chi-square statistic, Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) provided in Mplus (Muthén & Muthén, 2012). The model chi-square ($\chi 2$) statistic is the original fit index for the structural model (Wang & Wang, 2012), which assesses the overall fit and discrepancy between the sample and fitted covariance matrices. When $\chi 2$ is significant, the null hypothesis will be rejected such that the model fits the population and vice-visa. The degree of freedom (df) is the discrepancy between the number of pieces of available information less the number of estimated model parameters. The chi-square statistic is expressed by

$$\chi^2 = f_{ML}\,(N-1) \tag{1}$$

where fML = $(S, \Sigma^{\wedge})$ is the model-specific minimum fit function value, and N is the sample size. The cut-off for a good fit is p-value > 0.05

Comparative Fit Index (CFI): The CFI belongs to the group of incremental fit indices that compare the fit of the hypothesised model to the fit of a baseline model, which is the independent model. It demonstrates how much the hypothesised model fits better than the more parsimonious independence model. The CFI is a measure based on the noncentrality parameter d = ($\chi 2$ – df) where df represents the degree of freedom of the model. The formula for the CFI is

$$\text{CFI} = \frac{d_{null} - d_{specified}}{d_{null}} \tag{2}$$

where d_{null} and $d_{specified}$ are the noncentrality parameters for the null model and the specified model. The rule of thumb cutoff for the CFI is 0.90 but increased from 0.90 to 0.95 by Hu and Bletter (1999).

Tucker-Lewis Index (TLI): Another way to compare the lack of fit of a specified model to the lack of fit of the null model is the use of TLI. The TLI is also an incremental fit index, which does not guarantee a value from 0 to 1 and compares the fit of the target model to the fit of the independence model. The TLI is also known as the Non-Normed Fit Index (NNFI) and it is expressed as:

$$\text{TLI} = \frac{\left(\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{specified}}{df_{specified}}\right)}{\left(\frac{\chi^2_{null}}{df_{null}} - 1\right)} \tag{3}$$

where χ^2_{null} / df_{null} and $\chi^2_{specified}$ / $df_{specified}$ are ratios of χ^2 statistics to the degrees of freedom of the null model and the specified model, respectively. TLI has punishment for model complexity because the freer parameters, the smaller the $df_{specified}$, thus the larger $\chi^2_{specified}$ / $df_{specified,}$ leading to a smaller TLI model (Wang & Wang, 2012). A value greater than 0.95 has been regarded as the rule of thumb or cut-off criteria.

Root Mean Square Error of Approximation (RMSEA): The coefficient of the RMSEA is used to measure the approximate model fit. It is based on the non-centrality parameter:

$$\text{RMSEA} = \sqrt{\frac{(\chi_S^2 - df_S)}{df_S}} = \sqrt{\frac{(\chi_S^2/df_S) - 1}{N}} \quad (4)$$

where $(\chi_S^2 - df_S)$/N is the rescaled noncentrality parameter to adjust for sample size. It is understood that RMSEA values range from 0 to 0.10, where 0 indicates perfect fit, < 0.05 indicates close fit, 0.05-0.08 implies fair fit, 0.08-0.10 implies mediocre fit, and > 0.10 indicates poor fit (Browne & Cudeck, 1993; MacCallum et al., 1996; Byrne, 1998). A good model fit is defined as RMSEA ≤ 0.06 by Hu and Bentler (1999).

In view to assessing the dimensionality of the instrument, examinees' responses were subjected to the Stout test of essential unidimensionality (Stout, 1987), a nonparametric analysis using the Dimensionality Test (DIMTEST) package.

Results

Dimensionality Assessment

In psychological data involving subscales, one of the critical aspects that should be taken into account is the dimension of the data. A tenable assumption of unidimensionality needs to be made in any Item Response Theory (IRT) research context. There might be some degree of multidimensionality implied by an instrument with subscales. Tate (2002) states that when an instrument includes a subscale, there must be two aspects of validity that must be considered from a validity standpoint, namely the validity of the instrument's internal structure and the validity of the subscale's discriminant validity. Considering the assumption of unidimensionality, the first argument can be made. A dimensionality analysis should be performed before a DBF is assessed to ensure that the data are reasonably unidimensional (McCarty et al., 2007). In addition, dimensionality assessments can be useful in tests to determine whether or not the unidimensionality assumption used in the Unidimensional Item Response Theory (UIRT) has been strongly violated and may be used to measure whether or not this assumption has been at odds with the experimental results. Nevertheless, if evidence suggests that the unidimensionality of an item response theory is violated, then alternative methods can be used to find scores, such as those based on the Multidimensional Item Response Theory (MIRT). It is also possible to make predictions using the dimensionality assessments of different bundles of data as well as determine how those results can be compared with each other from different bundles of data.

Research Question One: *What is the dimensionality of the 2017 NECO Mathematics items?*

The dimensionality could either be unidimensional or multidimensional. However, the Stout test of essential unidimensionality is obtainable by dividing the items into two different groups. The first group of items consists of the Assessment Subtest (AT), which is designed in a way that is homogeneous with the rest of the group while also being dimensionally different from the remainder of the items in the group. There is a second group of items known as the Partitioning Subtest (PT). These are items that are not included in the AT. The grouping of items into two can be done by adopting either exploratory or confirmatory analysis but in this study, the exploratory analysis in DIMTEST was implemented.

Table 2

The Dimensionality of 2017 NECO Mathematics Items

TL	TGbar	T	p-value
33.04	13.68	16.89	0.00

The result of the test of Stout's essential unidimensionality (Table 2) was used to investigate the assumption of unidimensionality of the instrument that might form a secondary dimension. Two subtests, AT and PT, were divided into the test. A dominant trait is chosen as the item that measures the dominant trait and the AT in the most effective way. It seems that these items measure best when measured in a direction distinct from the direction of the PT items. It was decided to use the HCA/CCPROX clustering procedure to select the AT and the DETECT statistics in DIMTEST to perform the analysis. These items' cluster was tested to ascertain if it was dimensionally distinct from the secondary dimension of the test. A random sample of 30% of the examinees' responses was used to select the AT (items clustered in AT are 1, 9, 10, 11,12, 14, 16, 19, 21, 24, 28, 30, 31, 32, 33, 34, 35, 38, 39, 41, 43, 45, 46, 48, 49, and 50), and the remaining 70% of the examinees' responses were used as PT. The null and alternative hypotheses were tested using DIMTEST as proposed by Stout (1987).

H_0: $AT \cup PT$ satisfies essential a unidimensionality ($d = 1$)

H_i: $AT \cup PT$ fails to satisfy $d = 1$

Both the AT and PT assess a dimension that is dominant in the null hypothesis, while the items in the AT partition are best described by a dimension unique from the items in the PT partition. There was a violation of the essential unidimensionality assumption in the mathematics test items ($T = 16.87$, $p = 0.00$), resulting in the null hypothesis being rejected (Table 2). In addition, Table 2 presented the conclusion that the variance in the responses to the questions observed in the tests was attributable to multiple dimensions rather than one, which was the case previously. As a result of the implication of the above, there was a violation of the unidimensionality assumption involving the 2017 Mathematics items. This means that the 2017 NECO Mathematics item has a multifaceted aspect that must be considered. A further indication of the multidimensional nature of the 2017 NECO Mathematics items was provided by the T value, which was found to be statistically significant. It has been suggested by Furlow et al. (2009) that the use of UIRT models with multidimensional test data can violate or contradict the notion that all test items are equally dimensional and that there may be a potential hazard in estimating item and bundle parameters.

To address the research objectives, parameters were estimated with ESEM in Mplus 7.4 (Muthén & Muthén, 2012). Although cross-loading items are more visible and practicable with EFA, it is crystal clear that better techniques and approaches are more evident with CFA than with EFA. ESEM integrates the advantages of EFA and CFA into its technique. Thus, researchers such as (Ayanwale, 2022; Sass, 2011; Schmitt, 2011) argued that ESEM was a better and more efficient method to adjust for cross-factor loading instead of latent variables analysis, which assesses a measurement model of constructs through CFA. The model fit was established using chi-square ($\chi 2$), the CFI, TLI, and RMSEA.

Research Question Two: *Is there any statistically significant effect of DBF on 2017 NECO Mathematics with respect to sex?*

To answer this research question, the content analysis (Table 1) was developed as items were set into different bundles by implementing the confirmatory approach, and ESEM was adopted in analysing the data. The result is presented in Table 3.

Table 3

Differential Bundle Functioning of 2017 NECO Mathematics Items with Respect to Sex

Bundle	Estimate	S.E.	Est./S.E.	P-Value
1	0.113	0.006	18.833	0.000
2	0.121	0.007	17.286	0.000
3	0.093	0.008	11.625	0.000
4	0.102	0.007	14.571	0.000
5	0.087	0.007	12.429	0.000
6	0.088	0.008	11.000	0.000

Note: S.E. = Standard Error; Est. = Estimate

Table 4

Summary of Model Fit of ESEM

χ^2	df	RMSEA	CFI	TLI	p
169573.408	1749	0.043	0.964	0.958	0.0000

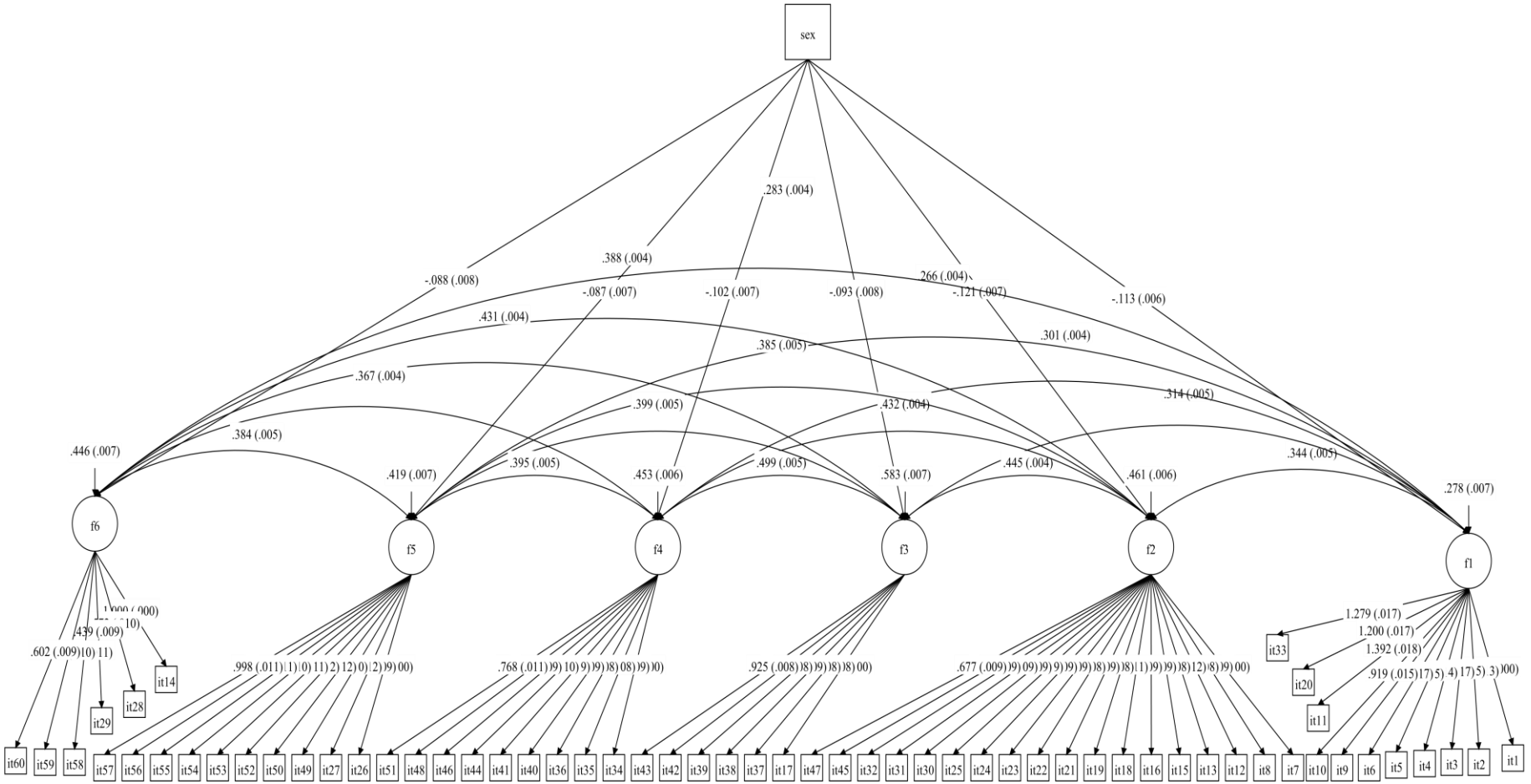
Note: RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index

The result (Table 3) showed that the items in the bundles possess construct validity as its fundamental factor functioned differentially with respect to examinees' sex. The result further displayed that all six bundles had a statistically significant effect on the examinees' sex with the underlying factor. the underlying factor of bundle 1, bundle 2, bundle 3, bundle 4, bundle 5, and bundle 6 are Z = 18.833, p = 0.000; Z = 17.286, p = 0.000; Z = 11.625, p = 0.000; Z = 14.571, p = 0.000; Z = 12.429, p = 0.000; Z = 11.000, p = 0.000 respectively. Table 4 further buttresses that the item bundles have construct validity as its fundamental factor functioned differentially with respect to examinees' sex. Also, the ESEM model was viable $\chi 2$ = 169573.408, df = 1749, p = 0.0000; RMSEA = 0.043 (90% C.I. = 0.043-0.043, probability of RMSEA $\leq$ 0.05 = 0.000), CFI = 0.964. TLI = 0.958. It also demonstrated the cross-loadings between all the items in the six bundles. However, the 2017 NECO Mathematics items at different bundles functioned differentially with respect to examinees' sex.

Figure 1 showed the model structure with estimated parameters of all the items and bundles in ESEM with respect to sex. It also demonstrated the cross-loadings between all the items in the six bundles. However, the 2017 NECO Mathematics items at various bundles functioned differentially with respect to examinees' sex.

Figure 1

Model Structure of Estimated Parameters with Respect to Sex



Research Question Three: *Is there any influence of DBF on 2017 NECO Mathematics items with respect to school type?*

To provide a valid answer to this research question, the content analysis (Table 1) was developed as items were grouped into bundles using the confirmatory approach based on items measuring the same construct, and ESEM was deployed in analysing the data. The results were presented in Table 5.

Table 5

Differential Bundle Functioning of 2017 NECO Mathematics Items with Respect to School Type

Variable	Estimate	S.E.	Est./S.E.	P-Value
Bundle 1	0.012	0.001	12.000	0.000
Bundle 2	0.009	0.001	9.000	0.000
Bundle 3	0.004	0.001	4.000	0.000
Bundle 4	0.006	0.001	6.000	0.000
Bundle 5	0.005	0.001	5.000	0.000
Bundle 6	0.005	0.001	5.000	0.000

Table 6

Summary of Model Fit Using ESEM

χ^2	Df	RMSEA	CFI	TLI	p
1235407.496	1830	0.043	0.964	0.958	0.0000

The result (Table 5) indicated that the item bundles pose construct validity as its fundamental factor functioned differentially with respect to the examinees' school type (public and private schools). The result also displayed that the bundles from bundles 1 to 6 had a statistically significant effect on the examinees' school type with the underlying factor. the underlying factor of bundle 1, bundle 2, bundle 3, bundle 4, bundle 5 and bundle 6 are Z = 12.000, p = 0.000; Z = 9.000, p = 0.000; Z = 4.000, p = 0.000; Z = 6.000, p = 0.000; Z = 5.000, p = 0.000 and Z = 5.000, p = 0.000 respectively. This implies (Table 6) that the item bundles have construct validity as its fundamental factor functioned differentially with respect to the examinees' school type. Also, the ESEM model (Table 6) was viable with $\chi 2$ = 1235407.496, df = 1830, p = 0.0000; RMSEA = 0.043 (90% C.I. = 0.043-0.043, probability of RMSEA $\leq$ 0.05 = 1.000), CFI = 0.964. TLI = 0.958. The data also had a good model fit as the CFI and TLI > 0.9. The differential performance noticed in the different bundles with respect to examinees' school type (public/private schools) may be attributed to the deficiency of an appropriate model for the test items, psychometric properties of the items not established, and lack of experience in the part of the item developer e.t.c.

Figure 2

Model Structure of Estimated Parameters with Respect to School Type

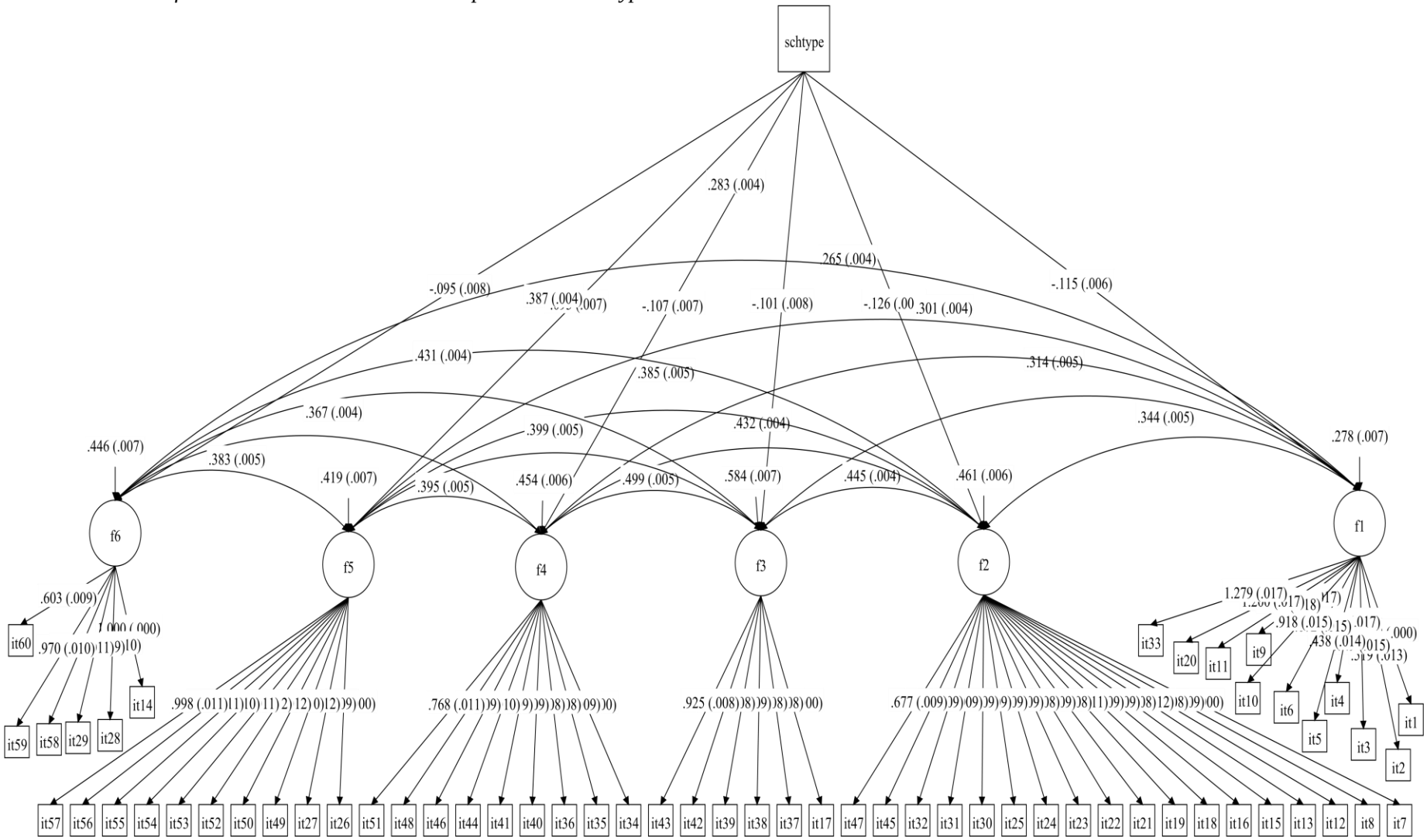


Figure 2 demonstrated the cross-loading of item bundles, and the bundles range from one to six. The cross-loading was achieved with the use of the ESEM. The differential performance noticed in the different bundles with respect to examinees' school type (public/private schools) may be attributed to the deficiency of an appropriate model for the test items, psychometric properties of the items not established, and lack of experience in the part of the item developer to mention but a few.

Discussion

This study aimed to ascertain the dimensionality of the 2017 NECO Mathematics items, determine the statistically significant effect of DBF of 2017 NECO Mathematics test items on examinees' sex, and finally, investigate the influence of DBF on 2017 NECO Mathematics items. However, the penultimate objective of the study was the use of the ESEM approach to determine the DBF of the 2017 NECO Mathematics test items. Furthermore, items were organised into various bundles using the confirmatory approach as postulated by Douglas et al. (1996). The 2017 NECO Mathematics test items had six bundles (Table 1) and each bundle was tested against the demographical variables (such as sex and school type) of the examinees. Based on the preliminary analysis of the study, it was revealed that the data is multidimensional and not unidimensional. Before DBF analysis, the dimensionality analysis must be established such that the data was reasonably unidimensional (Furlow et al., 2009). The application of UIRT models with multidimensionality data contradicts or violates the assumption of unidimensionality which invariably poses a statistically significant threat to bundle or item parameter estimates of examinees.

The dimensionality assessment of this study revealed the multidimensionality of the NECO Mathematics test items which was evident that more than one construct was measured. Similarly, suggestions have been made that tests like NECO comprise multiple-choice items, the different item types measure somewhat diverse traits, and therefore violate the IRT assumption of unidimensionality (Wainer & Thissen, 1993; Wainer et al., 1994). Also, it showed that more than one ability distribution is exhibited for an individual when the unidimensionality assumption of IRT has been compromised. When the assumption of unidimensionality has not been fulfilled, multidimensionality becomes the next alternative.

The second research question showed that there was a statistically significant effect of DBF on 2017 NECO Mathematics test items on the sex of examinees. This was in line with the study conducted by Boughton et al., (2000). They (Boughton et al., 2000) applied SIBTEST in understanding the differential performance of DBF on Mathematics and science achievement tests. They further revealed that male students consistently outperformed their female counterparts in Mathematics and Science. In addition, the result suggested that the model fit met the criteria postulated by Hu and Bentler (1999) that the CFI and TLI should be 0.90. ESEM uses either supplementary with CFA and it is an emerging technique used by researchers. Many studies (Marsh et al., 2020; Marsh et al., 2010; Marsh et al., 2009; Perry et al., 2015) conducted on ESEM have shown that ESEM is effective in the validation of a multidimensional measure like the 2017 NECO Mathematics test items. It was revealed in this study that the ESEM is a technique that is an appropriate substitute for CFA using Mathematics test items

The final research question also demonstrated a statistically significant influence of DBF on 2017 NECO Mathematics items with respect to the school type (public/private school) of the examinees. Walker et al. (2011) pointed out that the ability estimation bias can only be attributed to the DBF when a large number of items are showing whether focal groups of examinees perform differentially or not in a small way against that group of examinees or when a small number of items are showing differential performance against focal examinees in a large way. The existence of DBF in any standardised examination like NECO (which conducts a public examination that is used to adjudge whether a candidate is offered or denied admission into institutions of higher learning for Nigerian students) should be a cause of concern to the stakeholders in education. The essence is that the test scores obtained from the such national examination will be used to draw inferences about examinees' performance which will

invariably lead to overestimation or underestimation of examinees' ability thereby, leading to an erroneous judgment of the examinees' ability.

Based on the result, bundles feature items that share a common reading ability but may not share all cognitive tasks required for a correct response. The bundle of items associated with a mathematics test may be more difficult for an examinee who understands the term or does not understand the question. Also, the study finds that dependence within such bundles affects the distribution of items' responses in a predictable and testable manner. Although some small groups of items that share the same material exhibit excessive dependence, exam responses cannot be described as unidimensional. As a consequence, conventional IRT models can overestimate the standard error of measurement for exams with bundled items. Psychometric measurements for the six bundles are as follows: 0.51, 0.67, 0.68, 0.68, 0.65, and 0.44 respectively from bundles one to six. Some possible causes of those bundles of DBF in different groups of sex or school type might be the use of language structure of the items in the different bundle or when some items in the particular bundle focus on a certain area of interest like items relating to the sport. It is expected that male examinees might outperform their female counterparts in such items.

Conclusively, to ensure test fairness to all examinees, examination bodies like NECO should not only conduct DIF (DIF is not adequately proficient in detecting bias) but rather painstakingly apply ESEM which this study has shown to be effective in detecting DBF. Whenever DBF is detected, examination bodies are required to expunge or modify the item/item bundle (DBF) which can pose threat to the validity of an instrument which is the key focus of psychometricians.

Limitation to the study

The study was restricted to only NECO 2017 Mathematics test items, while further studies could be conducted on other subjects administered by NECO. A similar study should be conducted on various subjects of other public examining bodies such as West African Examination Council (WAEC), Joint Admission Matriculation Board (JAMB), and the National Business and Technical Examinations Board (NABTEB).

Declarations

Author Contribution: Oluwaseyi Aina Opesemowo: Conceptualization, methodology, analysis, writing & editing, visualization. Musa Adekunle Ayanwale: Methodology, analysis, writing & editing, visualization. Titilope Racheal Opesemowo: Conceptualization, methodology, writing & editing, visualization. Eyitayo Rufus Ifedayo Afolabi: Methodology, editing, and supervision.

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

Ethical Approval: Throughout this study, the researchers adhere to ethical principles. For the purpose of this study, secondary data were analyzed, which involved accessing pre-existing data that had been collected for research and anonymized. The study involved no direct participants, and no personal information was disclosed. This secondary data was used in accordance with the researchers' research plan and in a confidential and private manner. All relevant regulations, including institutional and national guidelines for data sharing and research ethics, were complied with by the researchers, who acknowledge and respect the efforts of those who collected and curated this dataset. Additionally, the researchers recognized that the data used represent individuals who might have contributed to their study, and they are committed to ensuring that their research will contribute to the advancement of psychometric knowledge. It is evident from the research conducted in this study that the researchers maintain the principles of respect for persons, beneficence, and justice and are committed to conducting their research in an ethical and responsible manner.

References

Abina, D. B. (2014). *Influence of teacher characteristics, availability and utilization of instructional materials on students' performance in mathematics* [Unpublished doctoral dissertation]. University of Ibadan.

Adegoke, B. A. (2011). Effect of direct teacher influence on dependent-prone students' learning outcomes in secondary school mathematics. *Electronic Journal of Research in Educational Psychology, 9*, 283-308.

Adeyegbe, S. O., & Oke, M. G. (2002). Science, technology and mathematics (STM) for sustainable development: The role of public examining bodies. *Proceedings of STAN annual conference* (pp. 144-147). Science Teachers Association of Nigeria.

Adeyemo, D. A. (2005). *Parental involvement interest in schooling and school environment as predictors of academic self-efficacy among senior secondary school students in Oyo State* [Unpublished doctoral dissertation]. University of Ibadan.

Adeyemo, E. O., & Opesemowo, O. A. (2020). Differential test let functioning (DTLF) in senior school certificate mathematics examination using multilevel measurement modelling. *Sumerianz Journal of Education, Linguistics and Literature, 3*(11), 249-253. https://doi.org/10.47752/sjell.311.249.253

Akale, M. A. G. (1997). The relationship between attitude and achievement among mathematics students in senior secondary school. *Journal of Science and Movement Education, 2*, 77-85.

Akinsola, M. K. (1994). *Comparative effects of mastery learning and enhanced mastery learning strategies on students' achievement and self-concept mathematics* [Unpublished doctoral dissertation]. University of Ibadan.

Aremu, O. A., & Sokan, B. O. (2003). *A multi-causal evaluation of academic performance of Nigerian learners: Issues and implication for national development*. Department of Guidance and counseling, University of Ibadan, Ibadan.

Asikhia, O. A. (2010). Students and teachers' perception of the causes of poor academic performance in Ogun state secondary schools: Implications for counseling for national development. *European Journal of Social sciences, 13*(2), 28-36.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

Awopeju, O. A., & Afolabi, E. R. I. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*, 263-284.

Ayanwale, M. A. (2019). *Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria* [Unpublished doctoral dissertation]. University of Ibadan.

Ayanwale, M. A. (2022). Performance of exploratory structural equation model (ESEM) in detecting differential item functioning. *EUREKA: Social and Humanities, 1*, 58–73. http://doi.org/10.21303/2504-5571.2022.002254

Black, S. (2001). Building blocks: How schools are designed and constructed affects how students learn. *American School Board Journal, 188*(10), 44-47.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Annual meeting of the Canadian Society for Studies in Education (CSSE), Edmonton, Alberta, Canada.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Sage.

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

Creswell, J. W. (2003). *Research design: Qualitative, quantitative and mixed methods approaches* (2nd ed.). Sage.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484. https://doi.org/10.1111/j.1745-3984.1996.tb00502.x

Finch, W. H. (2012). The MIMIC model as a tool for differential bundle functioning detection. *Applied Psychological Measurement, 36*(1), 40-59. https://doi.org/10.1177/0146621611432863

Furlow, C. F., Raiford, R. T., & Gagné, P. (2009). The impact of multidimensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement, 33*(6), 441-464. https://doi.org/10.1177/0146621609331959

Gierl, M. J., Tan, X., & Wang, C. (2005). *Identifying content and cognitive dimensions on the SAT®* (Report No. 2005-11). College Board.

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

Jennrich, R. I., & Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika, 31*(3), 313-323. https://doi.org/10.1007/BF02289465

Jöreskog, K. (1969). A general approach to confirmatory factor analysis. *Psychometrika, 34*, 183-202.

Latifi, S., Bulut, O., Gierl, M., Christie, T., & Jeeva, S. (2016). Differential performance on national exams: Evaluating item and bundle functioning methods using English, Mathematics, and Science Assessments. *SAGE Open, 6*(2), 1-14. https://doi.org/10.1177/2158244016653791

Lee, S., Bulut, O., & Suh, Y. (2016). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement, 77*(4), 545-569. https://doi.org/10.1177/0013164416651116

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. https://doi.org/10.1037/1082-989X.1.2.130

Marsh, H. W., Guo, J., Dicke, T., Parker, P. D., & Craven, R. G. (2020). Confirmatory factor analysis (CFA), exploratory structural equation modeling (ESEM), and set-ESEM: Optimal balance between goodness of fit and parsimony. *Multivariate Behavioural Research, 55*(1), 102-119. https://doi.org/10.1080/00273171.2019.1602503

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment, 22*(3), 471-491. https://doi.org/10.1037/a0019227

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*(1), 85-110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 439-476. https://doi.org/10.1080/10705510903008220

McCarty, F. A., Oshima, T. C., & Raju, N. S. (2007). Identifying possible sources of differential functioning using differential bundle functioning with polytomously scored data. *Applied Measurement in Education, 20*(2), 205-225. https://doi.org/10.1080/08957340701301660

Min, S., & He, L. (2020). Test fairness: Examining differential functioning of the reading comprehension section of the GSEEE in China. *Studies in Educational Evaluation, 64*, 100811. https://doi.org/10.1016/j.stueduc.2019.100811

Montoya, A. K., & Jeon, M. (2019). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement, 44*(2), 118-136. https://doi.org/10.1177/0146621619835496

Morin, A. J. S., & Maïano, C. (2011). Cross-validation of the short form of the physical self-inventory (PSI-S) using exploratory structural equation modeling (ESEM). *Psychology of Sport and Exercise, 12*(5), 540-554. https://doi.org/10.1016/j.psychsport.2011.04.003

Mucherah, W., Finch, W. H., & Keaikitse, S. (2012). Differential bundle functioning analysis of the self-description questionnaire self-concept scale for Kenyan female and male students using the MIMIC model. *International Journal of Testing, 12*(1), 78-99. https://doi.org/10.1080/15305058.2011.620724

Muthén, L., & Muthén, B. (2012). *Mplus user's guide* (7th ed.). Muthén and Muthén.

Ojimba, D. P. (2012). Strategies for teaching and sustaining mathematics as an indispensable tool for technological development in Nigeria. *Journal of Mathematical Sciences, 3*, 23-35.

Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing, 15*(4), 337-355. https://doi.org/10.1080/15305058.2015.1057639

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*(4), 353-369. https://doi.org/10.1207/s15324818ame1104_4

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12-21. https://doi.org/10.1080/1091367X.2014.952370

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347–363. http://doi.org/10.1177/0734282911406661

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*(4), 304–321. http://doi.org/10.1177/0734282911406653

Shealy, R., & 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. https://doi.org/10.1007/BF02294572

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617. https://doi.org/10.1007/BF02294821

Tata, U. S., Abba, A., & Abdullahi, M. S. (2014). The causes of poor performance in mathematics among public senior secondary school students in Azare Metropolis of Bauchi State, Nigeria. *IOSR Journal of Research & Method in Education, 4*, 32-40.

Tate, R. (2002). Test dimensionality. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). Lawrence Erlbaum.

Tsigilis, N., Gregoriadis, A., Grammatikopoulos, V., & Zachopoulou, E. (2018). Applying exploratory structural equation modeling to examine the student-teacher relationship scale in representative Greek sample. *Frontiers in Psychology, 9,* 733. https://www.frontiersin.org/article/10.3389/fpsyg.2018.00733

Umameh, M. A. (2011). *A survey of factors responsible for students' poor performance in mathematics in Senior Secondary School Certificate Examination (SSCE) in Idah Local Government Area of Kogi State, Nigeria* [Unpublished BSc(ED) thesis]. University of Benin.

Uwadie, I. (2012). *Federal government, teachers and parents battle students' under-performance*. Vanguard Newspaper. Retrieved September 23, 2022.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118. https://doi.org/10.1207/s15324818ame0602_1

Wainer, H., Wang, X.-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees choice? *Journal of Educational Measurement, 31*(3), 183-199. https://doi.org/10.1111/j.1745-3984.1994.tb00442.x

Walker, C. M., Zhang, B., Banks, K., & Cappaert, K. (2011). Establishing effect size guidelines for interpreting the results of differential bundling functioning analyses using SIBTEST. *Educational and Psychological Measurement, 72*(3), 415-434. https://doi.org/10.1177/0013164411422250

Wang, J., & Wang, X. (2012). *Structural equation modeling with Mplus methods and applications.* Wiley/Higher Education Press.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 19-32

Examining Measurement Invariance in Bayesian Item Response Theory Models: A Simulation Study

Merve AYVALLI* Hülya KELECİOĞLU**

Abstract

The aim of the study is to determine a measurement invariance cut-off point based on item parameter differences in Bayesian Item Response Theory Models. Within this scope, the Bayes factor is estimated for testing measurement invariance. For this purpose, a simulation study is conducted. The data were generated in the R software for each simulation condition under the one-parameter logistic model for 10 binary (1-0 scored) items. The invariance test was performed for various group sizes (n=500, 1000, 1500 and 2000) and difficulty parameters (d_k=0, d_k=0.1, d_k=0.3, d_k=0.5 and d_k=0.7). The Bayesian analyzes were performed on the WINBUGS using the codes written in the R. A Bayes factor that provides evidence for measurement invariance was calculated depending on the parameter differences. The Savage–Dickey density ratio, one of the MCMC sampling schemas, was used to calculate the Bayes factor. As a result, if the item parameter difference is d_k=0.3 and group sizes are 1500 or larger, the measurement invariance cannot be achieved. However, for small sample sizes (n=1000 or less) measurement invariance interpretation should be done carefully. When the d_k=0.5, there are invariant items only in n=500. According to Bayes factor test results, evidence has been produced when d_k=0.7 that measurement invariance cannot be achieved.

Keywords: Measurement invariance, bayesian IRT models, bayes factor, random item effects modelling

Introduction

The frequency of tests applied in education and psychology to measure latent variables such as cognitive and affective characteristics in groups having different characteristics has been progressively increasing. These kinds of testing applications often include a comparison among specific groups. Especially in the international large-scale assessments which aim to make comparisons against time or among different groups in terms of their mathematics, science, or reading skills as well as other psychological structures such as attitude, motivation, or anxiety (Davidov et al., 2014). In order to make meaningful comparisons among groups the measured latent variable must be the same in all subgroups. The measurement invariance is an important prerequisite for making comparisons between individuals or groups with varying demographic characteristics, such as different cultures, genders, or regions, to which the measurement tool is applied, by considering these differences. This is important to ensure the generalizability of the measured structure in different groups (Brown, 2006).

There are several methods for testing measurement invariance (Millsap, 2011). These can be examined in two different groups. One of these methods is the confirmatory factor analysis-based method. In the confirmatory factor analysis-based methods, the measurement invariance is examined by testing the similarity of measurement models between groups. One of the most important advantages of this method is that measurement invariance can be examined in all aspects such as factor loadings, intercepts,

* PhD. Student, Hacettepe University, Faculty of Education, Ankara-Türkiye, merveayvalli@gmail.com, ORCID ID: 0000-0002-7301-0096
** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Türkiye, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934
The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Merve AYVALLI.

To cite this article:
Ayvallı, M., & Kelecioğlu H. (2023). Examining measurement invariance in Bayesian Item Response Theory models: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 19-32. https://doi.org/10.21031/epod.1101457

Received: 11.04.2022
Accepted: 3.10.2022

residual variances, factor variances and covariance, and the latent-means. The Measurement invariance is tested by making comparisons among nested models (Meredith, 1993; Steenkamp and Baumgartner, 1998). The Higher levels of invariance require strict parameter equality and constraints between groups, which is difficult to meet in real applications.

The other group of methods includes the Item Response Theory (IRT) based methods. In the IRT-based methods, measurement invariance is tested by item bias methods used to evaluate the values of item-level observations in subgroups. Unlike the confirmatory factor analysis (CFA) based methods assuming a linear relationship at the item level, in the IRT based methods, a non-linear relationship is revealed between the latent structure and the item level scores. The Lord's $\chi 2$, Raju's area measures, Wald statistics (likelihood ratio test), Mantel-Hanzel procedure are among the IRT based methods (Millsap, 2011). However, all of these methods have some limitations such as the inability to provide evidence for the measurement invariance hypothesis and requiring to identify anchor items before the analysis (Verhagen, Levy, Millsap, and Fox, 2016).

Verhagen and Fox (2013) suggested a Bayes Factor on the basis of the variance of the item parameters between groups in to compare the measurement invariance hypothesis in nested and large groups such as countries and schools. The calculation of the Bayes Factor provides evidence both in favour and against the measurement invariance hypothesis, unlike the frequentist methods. In addition, anchor items are not needed for this method (Verhagen et al., 2016). However, this method is not convenient to compare a small number of groups.

Verhagen et al. (2016) proposed a different Bayesian factor that allows the comparison of a small number of groups and testing measurement invariance. The presented Bayesian measurement model is discussed and presented within the framework of a one-parameter logistic model.

In a test with i person (i=1, …, N) and k binary items (k=1, …, K), the probability of a correct response in the one-parameter logistic model (θ: ability parameter, bk: item difficulty parameter) as shown Equation 1 (Wright, 1977):

$$P(Y_{ik} = 1|\theta_i, b_k) = \frac{e^{(\theta_i - b_k)}}{1+e^{(\theta_i - b_k)}} \qquad (1)$$

In the Bayesian IRT models, item parameters' priors determine the alterations of item characteristics. In the random item effect models, all test items are considered as a random sample of the item population. The item parameters' priors for all items show standard normal distribution with a common mean and variance (Janssen, Tuerlinckx, Meulders & De Boeck, 2000; De Boeck, 2008).

$$b_k \sim N(b_0, \sigma_{b_k}^2)$$

The posterior distributions for each parameter are the functions of the combination of the average percent accuracy of that item in all groups and a prior distribution, the b_0 and $\sigma_{b_k}^2$. The standard normal prior is selected for the prior distributions of the person parameters (Fox, 2010). In the measurement invariance test, the Bayesian IRT model is considered within the scope of multiple groups IRT because multiple group comparisons are made. The multiple-group IRT models allow differences in test scores and item characteristics among groups (Bock & Zimowski, 1997). Thus, in the Bayesian IRT model, a measurement model is created by considering the variation of group-specific item parameters between groups, as well as the variation among the items. The probability of a person's correct response in multiple group IRT one-parameter logistic model is shown in Equation 2 (j group, i person, θ_{ij}= group-specific person parameter, $\tilde{b}_{kj}$= Group-specific item parameter):

$$P(Y_{ijk} = 1|\theta_{ij}, \tilde{b}_{kj}) = \frac{e^{(\theta_{ij} - \tilde{b}_{kj})}}{1+e^{(\theta_{ij} - \tilde{b}_{kj})}} \qquad (2)$$

Group-specific person parameter θ_{ij}, which is hierarchically modeled, shows a normal distribution around group mean μj.

$$\theta_{ij} \sim N(\mu_j, \sigma_j)$$

In the multiple-group IRT models, it is assumed that group-specific item parameters have a multilevel structure for modeling the measurement variance (Fox, 2010). The Group-specific deviations have normal distributions with a mean of zero and $\sigma^2_{b_k}$ for all items. This variance component defines the variability of item functions between groups. When this variance is zero, the item is considered to be invariant because there is no variability. In addition, if a measurement invariance study is desired between a small number or fixed groups, it is more useful to use the fixed group model instead of the random group model, since it will be difficult to estimate the random item effects variance.

Verhagen et al. (2016) introduced a model that can test measurement invariance between two groups with Bayes factor. In this model, the group-specific item parameters are estimated separately for different groups. The item parameters are independent, and they do not provide information about each other. In such a case, a possible prior distribution for the group mean is the normal prior distribution with a large variance. A multivariate normal model is applied to group-specific item characteristics. In addition, covariance matrices which are based on the correlation between the item parameters of different groups are used.

The group-specific item parameters defined in the model are shown in Equation 3 (μ_j: mean of the item difficulty in group j, e_{kj}: error term):

$$\tilde{b}_{kj} = \mu_j + e_{kj} \qquad (3)$$

In the model, μ_j equals zero and e_{kj} equals the amount of deviation from the average item difficulty in the group.

These deviations are assumed to show multivariate normal distribution with covariance (Σb) for item difficulties consisting of item parameter variances for each group. The variance of item parameters may vary by group. This means that the variance in the item difficulty parameters of one group is higher than the other. Since the group-specific item difficulties are estimated independently for each group, the measurement invariance can be directly estimated based on the differences between the item difficulty parameters. The difference between the difficulty parameters among the two groups is shown in Equation 4 (k. item, groups j and j', j<j'):

$$d_{kjj'} = b_{kj} - b_{kj'} \qquad (4)$$

In the measurement invariance test for the two groups, the hypotheses are established based on the difference between the item difficulty parameters.

$$H_0 : d_k=0$$
$$H_1 : d_k\neq0$$

The Bayes factor which uses Bayesian hypothesis testing is very advantageous in that it provides direct information about items' measurement invariance in a whole test (Jeffreys, 1961). Additionally, it does not require items that have been proven measurement invariance before (Verhagen et. al., 2016). Furthermore, unlike frequentist statistics, it gives evidence for both H_0 and H_1.

When only evidence for H_0 is given without using the evidence given for H_1, it leads to exaggerated results against H_0 hypothesis that only the evidence for the null hypothesis is considered, especially in low-power studies (Rouder et al., 2009; Wagenmakers et al., 2017). In addition, providing evidence for both hypothesis tests is advantageous in terms of giving information about which hypothesis should be preferred.

In the measurement invariance test, the marginal likelihood of the H_1 hypothesis is weighted by the prior probability of the average likelihood on all possible values of the alternative hypothesis. This average likelihood value is equal to the integral of the likelihood function weighted by the prior density function of the parameters in the hypothesis. The Bayes factor includes the marginal likelihood ratio for both the null hypothesis and alternative hypothesis results.

The Bayes factor that provides relative evidence for the hypotheses tested is as in Equation 5 ($p_1(d_k)$: H_1 under Cauchy prior distribution)

$$BF_{01} = \frac{P(Y|H_0)}{P(Y|H_1)} = \frac{P(Y|d_k = 0)}{\int P(Y|d_k)p_1(d_k)dd_k} \qquad (5)$$

The increase in cross-cultural testing practices also increases the importance of measurement invariance. Traditional methods based on confirmatory factor analysis, which are frequently used in determining measurement invariance, require the comparison of different model fits. In addition, these methods are time-consuming as each of these models is expected to be fitted (White, 2000). Furthermore, additional restrictions are needed in the definition of these models (Reise, Widaman & Pugh, 1993).

Other invariance tests require at least one anchor item of which invariance has already been proven. The methods used to select an item that is invariant for many groups (Langer, 2008) lead to biased estimations if the item contains a certain level of bias (Woods, Cai, & Wang, 2012). Considering all these situations, more practical methods are needed to evaluate measurement invariance, especially in large-scale tests and cross-cultural studies. Unlike the frequentist methods, multiple hypotheses (H_0 and H_1) can be tested simultaneously with the Bayesian method that is used in the study. Thus, all items in the test can be evaluated simultaneously and measurement invariance estimation can be made directly. From both practical and theoretical perspectives, it is thought that the research results will be significant. It will be possible to have an idea about the measurement invariance based on the difference between item difficulty parameters. In addition, it is important that these cut-off points, which are limited by the conditions in the study, form an idea for both frequentist and Bayesian measurement invariance studies.

Purpose of the Study

The aim of the study is to determine a measurement invariance cut-off point based on item parameter differences. In accordance with this purpose, the Bayes factor which estimates invariance directly is used within the scope of Bayesian IRT models. The invariance test was performed when the difficulty parameter differences (d_k) are 0.0, 0.1, 0.3, 0.5, 0.7 at group sizes are 500, 1000, 1500, and 2000.

Method

The Simulation Conditions and Data Generation

The Data was generated in the R software (R Core Team, 2018) when evaluating measurement invariance using the Bayes factor. For each group, the sum of item thresholds (b_{kj}) and reference group's ability parameter ($\mu_{\theta j}$) are assumed to be zero. The sample sizes in groups were equally determined to be 500, 1000, 1500, and 2000. The previous studies suggested that the minimum sample size should be 500 for unbiased parameter estimation (Thompson, 2018, Asparouhov & Muthén, 2014; De Boeck, 2008; Stark et al., 2006). It has been found that by increasing the group size from 500 to 1,000 the Type I error rate decreased, but there was no significant difference in the Type I error when increasing the group size from 1,000 to 2,000 (Finch, 2016). And it was determined that the Bayes factor performed well in group sizes of 500 and more (Verhagen et al., 2016). Thompson (2018) noticed that Bayes Factor for measurement invariance has got higher power rate with larger sample sizes and suggested using at least 500 as a sample size. In addition, considering the real test applications such as PISA, TIMMS, and PIRLS, it is known that the minimum sample sizes are usually 500 and more. Based on these findings and real data applications the sample sizes in groups were equally determined to be 500, 1000, 1500, and 2000 in the study. The data were generated for each simulation condition under the one-parameter logistic model for 10 binary (1-0 scored) items. The difference between the difficulty parameters of the groups was determined as d_k=0, d_k=0.1, d_k=0.3, d_k=0.5 and d_k=0.7. d_k= 0.0 (there is no difference between item difficulty parameters) were considered as invariant items and the difference between the parameters gradually increased (Verhagen et al., 2016). Harwell et al. (1996) stated that 100 or fewer replications would have sufficient power in simulation studies and recommended at least 25 replications. In the current study, 100 replications were applied for each condition. The analyses were carried out for each data set. Item difficulty parameter values for each condition can be seen in Annex-1.

Data Analysis

Bayesian analyzes were performed on the WINBUGS using the codes written in the R. For the difference between the difficulty parameters of each item, a Bayes factor was created to provide evidence for the measurement invariance. In hypothesis testing, the ratio of the density of the null hypothesis under the prior and posterior distributions affects the Bayes factor test results. The Bayes factor test results depend on priors selected for the parameters to be evaluated. The priors can be selected based on the assumptions accepted for the parameter values. Since the Bayes factor is more likely to support measurement invariance when multivariate Cauchy prior is used. The difference between group-specific item parameters is equally distributed under the multivariate Cauchy prior. Thus, the analyses were performed using the multivariate Cauchy prior (Verhagen et al., 2016, Thompson, 2018). The Savage–Dickey density ratio, one of the MCMC sampling schemas, was used to calculate the Bayes factor.

This method is applied in nested models, and the calculation of the Bayesian factor for the parameter under test requires high posterior and prior distribution. Especially in complex models, such as nested structures, this method can be used for invariance testing. In this model, the null hypothesis is the hypothesis that the value of the parameter of interest is fixed, and the alternative hypothesis is the hypothesis that this parameter is released. Therefore, the null hypothesis is nested under the alternative hypothesis (Wagenmakers, Lodewyckx, Kuriyal, and Grasman, 2010). The difference between item difficulty parameters for any of the two groups is defined as:

$$d_k = b_{k1} - b_{k2} = 0$$

The Bayes factor reduces the H_0 to the prior and posterior distributions of the difference between parameters in the H_1, when evaluating the relative support of the H_0= d_k=0 according to the H_1= $d_k \neq 0$ hypothesis. In a simpler expression, it is obtained from the alternative hypothesis by setting it to H_0= 0.

$$BF_{01}=\frac{p_1\,(d_k=0\mid H_1,Y)}{p_1\,(d_k=0\mid H_1)}$$

Thus, the Bayes factor produces more evidence for the null hypothesis than for the alternative hypothesis (Verhagen et al., 2016).

The Bayes factor defines a relative estimation performance for the H_0 and H_1. In other words, it specifies a relative measure of the prediction quality of the hypothesis. For instance, if BF_{01}=5, it means that the data is 5 times more likely to be under H_0 than under H_1. However, the fact that the Bayes factor favors H_0 does not mean that H_0 predicts the data better (van Doorn, van den Bergh, Böhm et al., 2021). According to Jeffreys (1961), the Bayes factors between 1 and 3 produce equal evidence for the null hypothesis and the alternative hypothesis, and these values are accepted as weak evidence. A Bayes factor between 3 and 10 is considered sufficient evidence for the H_0 hypothesis. If the Bayes factor is greater than 10, it is accepted as strong evidence for the H_0 hypothesis. When the Bayes factor is between 0.33 and 0.10, it is accepted as sufficient evidence for the alternative hypothesis, and when it is less than 0.1, it is accepted as strong evidence for the H_1. In the current study, the cut-off point for the Bayes factor was determined as 3 if the invariance holds, and 0.33 if it does not hold. To complete the MCMC processes efficiently, the analysis was carried out with 3000 iterations with a 300 burn-in period.

Findings

The measurement invariance was tested when there is no difference between the difficulty parameters. The Results are shown in Table 1.

Table 1

The Bayes Factor Results for d_k=0.0

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.0	**Item_1**	4.59773	6.30162	10.37240	18.34086
	Item_2	10.85618	10.0564	7.82427	16.83027
	Item_3	10.67626	11.04705	7.63375	11.0673
	Item_4	11.57481	9.56177	16.39697	15.62178
	Item_5	6.24642	11.08648	21.14449	14.84418
	Item_6	5.50773	6.36254	10.24370	19. 08634
	Item_7	10.97618	11.0004	8.84272	17.80372
	Item_8	12.67626	10.99905	7. 75363	12.00662
	Item_9	15.57481	9.59548	17.12697	16.16278
	Item_10	6.43642	13.08868	22.00443	15.73325

According to the results for d_k=0.0, in all sample sizes, it is seen that the BF_{01} values of the item parameters are greater than the cut-off point of 3. When the group sizes are 500, 1000, and 1500, the BF_{01} values of 4 items were greater than 3, and the BF_{01} values of 6 items were greater than 10. Since the group size is 2000, it is seen that BF_{01} values for all items are greater than 10, which provides strong evidence for the measurement invariance. The Bayes factor test results can be seen in Table 2 when the difference between parameters is dk=0.1 for each sample size.

Table 2

Bayes Factor Results for d_k=0.1

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.1	**Item_1**	9.53619	13.53914	3.94931	9.53863
	Item_2	3.32114	10.49564	10.93055	4.85995
	Item_3	8.24058	8.81567	20.05434	6.65487
	Item_4	11.84632	9.00784	12.1836	16.94075
	Item_5	7.99764	12.38298	9.74235	6.89683
	Item_6	9.78869	15.91453	3.00491	9.63375
	Item_7	5.32114	10.56449	11.04056	4.99502
	Item_8	8.42058	7.99815	21.45434	7.48765
	Item_9	13.87632	9.78400	13.18360	17.93081
	Item_10	8.19765	12.29809	9. 35945	7. 96828

The results in Table 2 show that when the parameter differences are d_k=0.1, it is seen that BF_{01} values are greater than 3 in all sample sizes. It is seen that measurement invariance is obtained for all items. In cases where the difference between the parameters is d_k=0.3, the Bayes factor results calculated based on the difference between the item difficulty parameters are given in Table 3.

When the difference between item difficulties is 0.3, there are invariant items only n=500 and n=1000. The Bayes factor results for 6 items are invariant (n=500). However, there are 4 items producing equal evidence for both the null hypothesis and the alternative hypothesis. Therefore, those items cannot be interpreted as invariant. It was seen that if the group size was 1000 and the difference between the parameters was 0.3, the measurement invariance was obtained in 4 items. Invariance interpretation for 2 items cannot be made because of the equal evidence for the hypotheses (H_0 and H_1).

Table 3

Bayes Factor Results for d_k=0.3

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.3	**Item_1**	7.19235	10.3738	2.61409	0.02789
	Item_2	12.04128	1.12155	0.29342	0.05877
	Item_3	8.74905	0.20997	0.32324	0.09858
	Item_4	1.92266	4.18721	0.01816	0.27239
	Item_5	1.32949	0.0207	0.01307	0.13994
	Item_6	8.19235	13.3738	2. 96104	0.02789
	Item_7	11.94128	1.15585	0.19388	0.05877
	Item_8	7.14901	0.19997	0.23564	0.09858
	Item_9	0.99266	4.72118	0.02817	0.27239
	Item_10	2.31742	0.20700	0.01509	0.13994

When the group size was 1500, there is not an invariant item. According to the Bayes factor results for n=1500, there are 2 items having equal evidence for both H_0 and H_1. That can be considered weak evidence for both hypotheses. It can be said that the values of BF_{01} for the remaining 8 items are less than 0.33, and the items are not invariant. The Bayes factor results are less than 0.33 for n=2000 which means none of the items were invariant. Table 4 shows the Bayes factor results calculated based on the difference between the item difficulty parameter which is d_k=0.5.

Table 4

Bayes Factor Results for d_k=0.5

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
d_k=0.5	**Item_1**	0.15524	0.06085	0.00168	0.00016
	Item_2	0.18284	0.00891	0.01364	0.03911
	Item_3	4.09778	0.00085	0.00116	0.00415
	Item_4	0.35232	0.02586	0.06630	0.00147
	Item_5	4.31605	0.03253	0.00393	0.00135
	Item_6	0.23245	0.06857	0.01368	0.00012

Table 4

Bayes Factor Results for d_k=0.5 (Continued)

	Item_7	0.17294	0.00981	0.02264	0.01368
	Item_8	4.00038	0.00508	0.01015	0.00307
d_k=0.5	**Item_9**	0.29852	0.01258	0.06730	0.00112
	Item_10	3.93167	0.03534	0.00298	0.00129

As in Table 4, there are invariant items only for a group size of 500. In other group sizes, there are no items with a Bayes factor value greater than 3. When n=500, the 4 items are invariant, but the remaining items are not.

In all remaining group sizes (n=1000, n=1500, and n=2000), all BF_{01} values of the items are less than 0.10, which provides strong evidence in favor of the H_1 hypothesis. In Table 5, the Bayes factor results are shown for the cases where the difference between the difficulty parameters is d_k=0.7.

Table 5

Bayes Factor Results for d_k=0.7

		BF_{01}			
		N=1000 (500 per group)	***N=2000 (1000 per group)***	***N=3000 (1500 per group)***	***N=4000 (2000 per group)***
	Item_1	0.32990	0.01297	0.00030	0.00003
	Item_2	0.00047	0.00172	0.00156	0.00020
	Item_3	0.00328	0.00136	0.00077	0.00113
	Item_4	0.00232	0.02296	0.00175	0.00008
	Item_5	0.06634	0.00223	0.00026	0.00026
d_k=0.7	**Item_6**	0.26890	0.02129	0.00140	0.00002
	Item_7	0.00007	0.00147	0.00185	0.00009
	Item_8	0.00285	0.00163	0.00113	0.00126
	Item_9	0.00292	0.03295	0.00156	0.00017
	Item_10	0.07654	0.02019	0.00102	0.00032

According to the Bayes factor test results shown in Table 5, there is no evidence for measurement invariance in all group sizes. BF_{01}values for 2 items are less than 0.33 only when the group size is 500, in all other cases, it was determined that the BF_{01} value was less than 0.10 and produced strong evidence in favor of the H_1 hypothesis.

Conclusion

In the study, it was aimed to determine a cut-off point for measurement invariance based on the difference between parameters in different sample sizes and in cases where item parameters differed between groups with the Bayesian IRT model.

According to simulation results;

1. As predicted, it was determined that measurement invariance was achieved in all sample sizes when there was no difference between the difficulty parameters of the groups.
2. When the difference between item difficulty parameters is d_k=0.1, all items are invariant for all sample sizes.
3. Bayes factor results for d_k=0.3 shows that only a few items are invariant if the group sizes are 500 and 1000. It has been found that the number of invariant items decreases as the group size increases. When the group sizes are 1500 and 2000, the Bayes factor test results provide evidence for only the alternative hypothesis. Thus, there is no invariant item for these sample sizes.
4. When the difference between item difficulty parameters is d_k=0.5, there are invariant items only in n=500. Bayes factor results provide strong evidence in favor of H_1 for n=1000, n=1500, and n=2000.
5. There is no evidence in favor of invariant items when d_k=0.7 for all sample sizes.

When the results are evaluated, it is seen that no invariant item was detected independent of group size when the difference between the item difficulty parameters is d_k=0.7. In this situation, it is possible to state that if the item difficulty parameters difference between groups is 0.7, measurement invariance does not hold.

In cases where the d_k=0.5 and the sample size is 1000 or larger, it can be said that measurement invariance cannot be achieved. However, if the group size is n=500 or smaller, it is not possible to evaluate the invariance only based on the item parameter differences. For these sample sizes, it is recommended to perform a measurement invariance test at the item level.

For d_k=0.3 and n=2000 or larger, the measurement invariance is not achieved. But, for n= 1500 or smaller, it is not correct to make a final decision for measurement invariance based solely on the differences among the item parameters. To make a decision on the measurement invariance, the invariance test must be performed.

If there is no difference between difficulty parameters or d_k=0.1, it is possible to say that measurement invariance is achieved in all group sizes.

There are many studies related to the Bayesian approximate invariance and alignment optimization method to determine Bayesian measurement invariance. On the other hand, the studies are limited to investigating measurement invariance with the Bayes factor. In the literature, Verhagen (2013) stated that the Bayes factor performs well in detecting the measurement invariance when the difference between the item difficulty parameters is large (d_k>0.5). If there is a smaller difference (d_k=0.1 or d_k=0.3), the Bayes factor could not decide on invariance for most items. Also, Verhagen et al. (2016) have shown that the Bayes factor is a valid method for determining invariance when the difference between item difficulty parameters is 0.5 or more. It will be easier to detect measurement invariance with respect to the cut-off point, especially when group sizes are large. Thompson (2018) has shown that the Bayes Factor distinguishes invariant and non-invariant items.

In conclusion, providing measurement invariance is a prerequisite for meaningful comparisons, especially when comparisons between groups are required (Horn ve McArdle, 1992). The current study revealed that it is possible to have an idea about the measurement invariance based on item difficulty parameter differences and it creates a practical framework for doing meaningful comparisons across groups. Defining which items are most likely to be invariant with the difference between item difficulty parameters, provides pragmatic information for measurement invariance and differential item

functioning studies. The cut-off points presented in the study can be used in applications that are compatible with the conditions described in the research. In addition, as a secondary outcome, when an anchor item needs to be determined, the item selections can be made by considering the cut-off points d_k=0.0 and d_k=0.1

The current research is limited to the simulation conditions explained in detail in the method section, and binary items. Studies on polytomous items, unequal sample sizes, different simulation conditions, and real data applications can be conducted in future studies. Furthermore, the study focused on only the Bayes factor for detecting measurement invariance. In future applications, studies can be performed using not only the Bayes factor but also other Bayesian criteria such as Deviance Information Criteria (DIC).

Declarations

Author Contribution: Merve Ayvallı-Conceptualization, methodology, analysis, writing & editing, visualization. Hülya Kelecioğlu-Conceptualization, methodology, writing-review & editing, supervision.

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

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

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. https://doi.org/10.1080/10705511.2014.919210

Bock, R. D., & Zimowski, M. F. (1997). The multiple groups IRT. In Wim J. van der Linden, & Ronald K. Hambleton (Eds.), *Handbook of modern item response theory*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2691-6_25

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Newyork: Guilford Publications.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross -national research. *Annual Review of Sociology*, *40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*(4), 533-559. https://doi.org/10.1007/s11336-008-9092-x

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied measurement in Education*, *29*(1), 30-45. https://doi.org/10.1080/08957347.2015.1102916

Fox, J. P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-0742-4

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*(2), 101-125. https://doi.org/10.1177/014662169602000

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144. https://doi.org/10.1080/03610739208253916

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical irt model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306. https://doi.org/10.3102/10769986025003285

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Langer, M. M. (2008). A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation (Doctoral dissertation, The University of North Carolina at Chapel Hill).

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

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge. https://doi.org/10.4324/9780203821961

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-90. https://doi.org/10.1086/209528

Thompson, Y. T. (2018). Bayesian and Frequentist Approaches for Factorial Invariance Test (Doctoral dissertation, University of Oklahoma).

Van Doorn, J., van Den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š. L., Marsman, A., Matzke, M., Gupa, D., R, A., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28(3), 813–826. https://doi.org/10.3758/s13423-020-01798-5

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383-401. https://doi.org/10.1111/j.2044-8317.2012.02059.x

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*, *72*, 171-182. https://doi.org/10.1016/j.jmp.2015.06.005

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158-189. https://doi.org/10.1016/j.cogpsych.2009.12.001

Wagenmakers, E. J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.). *Psychological science under scrutiny: Recent challenges and proposed solutions*, (pp. 123-138). https://doi.org/10.1002/9781119095910.ch8

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097-1126. https://doi.org/10.1111/1468-0262.00152

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*, 532–547. https://doi.org/10.1177/0013164412464875

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of educational measurement*, *14*(2) 97-116. https://www.jstor.org/stable/1434010

Annex 1

Difference between parameters	Items	Group 1 N= 500	Group 2 N= 500	Group 1 N= 1000	Group 2 N= 1000	Group 1 N= 1500	Group 2 N= 1500	Group 1 N= 2000	Group 2 N= 2000
d_k=0.0	**Item_1**	-1.51524	-1.51524	1.35099	1.35099	-0.14897	-0.14897	0.61867	0.61867
	Item_2	1.15225	1.15225	0.15923	0.15923	0.25329	0.25329	-1.20775	-1.20775
	Item_3	-0.58729	-0.58729	-1.04742	-1.04742	-0.13534	-0.13534	-0.08178	-0.08178
	Item_4	-1.17103	-1.17103	-2.31015	-2.31015	1.17636	1.17636	-0.95174	-0.95174
	Item_5	-0.19137	-0.19137	1.41825	1.41825	-0.44912	-0.44912	-1.66571	-1.66571
	Item_6	1.00173	1.00173	-0.8556	-0.8556	-0.85863	-0.85863	1.11071	1.11071
	Item_7	0.92033	0.92033	-0.22294	-0.22294	0.10577	0.10577	-0.18306	-0.18306
	Item_8	-0.19002	-0.19002	-0.1573	-0.1573	-0.77499	-0.77499	-0.09381	-0.09381
	Item_9	0.13939	0.13939	1.84483	1.84483	0.7638	0.7638	1.10559	1.10559
	Item_10	0.44124	0.44124	-0.17991	-0.17991	0.06784	0.06784	1.34887	1.34887
d_k=0.1	**Item_1**	0.11130	0.21130	1.47472	1.57472	-1.14341	-1.04341	-1.25033	-1.15033
	Item_2	-1.42535	-1.52535	-0.72646	-0.82646	0.64185	0.54185	-0.98874	-1.08874
	Item_3	1.0506	1.1506	1.28423	1.38423	-0.23042	-0.13042	0.46737	0.56737
	Item_4	0.61857	0.51857	0.72544	0.62544	-0.14181	-0.24181	0.28588	0.18588
	Item_5	-1.17563	-1.07563	-1.6525	-1.5525	0.33662	0.43662	-0.4916	-0.3916
	Item_6	0.85522	0.75522	-0.6442	-0.7442	1.94682	1.84682	1.52708	1.42708
	Item_7	-0.72103	-0.62103	-0.2051	-0.1051	0.04659	0.14659	-0.2995	-0.1995
	Item_8	0.67932	0.57932	1.21579	1.11579	-0.46611	-0.56611	-1.12467	-1.22467
	Item_9	-1.13605	-1.03605	-0.84939	-0.74939	0.98127	1.08127	0.72206	0.82206
	Item_10	1.14304	1.04304	-0.62251	-0.72251	-1.97141	-2.07141	1.15243	1.05243
d_k=0.3	**Item_1**	-0.83200	-0.53200	1.39128	1.69128	-0.24920	0.05080	0.31251	0.61251
	Item_2	2.23221	1.93221	-1.75713	-2.05713	-0.48799	-0.78799	-0.34488	-0.64488
	Item_3	1.19172	1.49172	-0.58153	-0.28153	0.23554	0.53554	0.28424	0.58424
	Item_4	-0.15327	-0.45327	1.3534	1.0534	0.28922	-0.01078	0.39478	0.09478
	Item_5	-0.85613	-0.55613	-0.49663	-0.19663	-1.10461	-0.80461	-1.6253	-1.3253
	Item_6	-1.24592	-1.54592	1.27965	0.97965	-0.53536	-0.83536	0.34952	0.04952
	Item_7	0.30058	0.60058	-0.97319	-0.67319	0.65091	0.95091	1.91602	2.21602

Annex 1 (Continued)

d_k=0.3	**Item_8**	-0.37292	-0.67292	-0.18041	-0.48041	0.55902	0.25902	-0.05588	-0.35588
	Item_9	-0.49208	-0.19208	-0.22857	0.07143	1.61207	1.91207	-1.14867	-0.84867
	Item_10	0.22780	-0.07220	0.19312	-0.10688	-0.96959	-1.26959	-0.08233	-0.38233
d_k=0.5	**Item_1**	1.01674	0.51674	0.69355	0.19355	0.38955	-0.11045	0.91923	0.41923
	Item_2	-0.78838	-0.28838	1.6497	2.1497	-0.76968	-0.26968	-0.24888	0.25112
	Item_3	-0.08844	-0.58844	-0.18656	-0.68656	1.59203	1.09203	-1.80859	-2.30859
	Item_4	-0.82591	-0.32591	0.92077	1.42077	0.83879	1.33879	-0.5192	-0.0192
	Item_5	0.88156	0.38156	0.36453	-0.13547	1.4836	0.9836	3.13911	2.63911
	Item_6	0.4986	0.9986	-0.801	-0.301	0.43809	0.93809	-1.70057	-1.20057
	Item_7	-0.47966	-0.97966	-0.60329	-1.10329	-0.74047	-1.24047	0.49179	-0.00821
	Item_8	-0.3249	0.1751	0.09697	0.59697	-1.30876	-0.80876	-0.14801	0.35199
	Item_9	-1.01617	-1.51617	-0.97367	-1.47367	-0.76574	-1.26574	0.23068	-0.26932
	Item_10	1.12655	1.62655	-1.16101	-0.66101	-1.15741	-0.65741	-0.35558	0.14442
d_k=0.7	**Item_1**	-0.02227	-0.72227	-0.06218	-0.76218	0.99252	0.29252	1.78028	1.08028
	Item_2	0.29443	0.99443	2.12972	2.82972	0.43886	1.13886	0.3036	1.0036
	Item_3	-0.88435	-1.58435	-0.77638	-1.47638	0.32672	-0.37328	0.74875	0.04875
	Item_4	0.98951	1.68951	0.12361	0.82361	-0.37344	0.32656	-0.46008	0.23992
	Item_5	0.00746	-0.69254	-0.58109	-1.28109	0.21024	-0.48976	-0.46803	-1.16803
	Item_6	-1.44076	-0.74076	-0.1527	0.5473	0.35615	1.05615	-1.92719	-1.22719
	Item_7	-0.21819	-0.91819	-1.40278	-2.10278	-0.45817	-1.15817	0.59344	-0.10656
	Item_8	2.1256	2.8256	0.84659	1.54659	-1.23148	-0.53148	0.20459	0.90459
	Item_9	-0.54605	-1.24605	0.42517	-0.27483	-0.31452	-1.01452	0.34729	-0.35271
	Item_10	-0.30538	0.39462	-0.54997	0.15003	0.05311	0.75311	-1.12264	-0.42264

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 33-46

The Effects of Different Item Selection Methods on Test Information and Test Efficiency in Computer Adaptive Testing

Merve ŞAHİN KÜRŞAD*

Abstract

The purpose of this study is to examine the effect of different item selection methods on test information function (TIF) and test efficiency in computer adaptive testing (CAT). TIF indicates the quantity of information the test has produced. Test efficiency resembles the amount of information from each item, and more efficient tests are produced from the smallest number of good-quality items. The study was conducted with simulated data, and the constants of the study are sample size, ability parameter distribution, item pool size, model of item response theory (IRT) and distribution of item parameters, ability estimation method, starting rule, item exposure control and stopping rule. The item selection methods, which are the independent variables of this study, are the interval information criterion, efficiency balanced information, matching -b value, Kullback-Leibler information, maximum fisher information, likelihood-weighted information, and random selection. In the comparison of these methods, the best performance in the aspect of TIF is provided by the maximum fisher information method. In terms of test efficiency, the performances of the methods were similar, except for the random selection method, which had the worst performance in terms of both TIF and test efficiency.

Keywords: Computer adaptive testing, test information function, test efficiency

Introduction

In the field of education, where individual differences are important, the use of individually adjusted tests, which are also called “tailored” or “adaptive” tests, has been increasing recently. As computerized adaptive tests (CAT) are individualized tests, each individual encounters different items according to their available ability, and their ability is recalculated after each item application. Therefore, everyone receives different tests (Eggen, 2004; Sulak, 2013).

In CAT applications, each individual is presented with different items according to their estimated ability level. In Item Response Theory (IRT) based CAT applications, individuals’ abilities and item difficulties are placed on the same scale, and their likelihood of answering correctly in the relevant ability is calculated with 50 per cent probability (Lord, 1980). This makes CAT applications preferable in terms of time and cost compared to traditional paper-pencil tests, as CAT applications conclude the test with fewer items and allow for as valid and reliable tests as paper-pencil test applications (Çıkrıkçı-Demirtaşlı, 1999; Kaptan, 1993; Wainer, 1993; Weiss & Kingsbury, 1984).

Item response theory and its models are important factors for CAT applications because, to match and evaluate the item and ability parameters, IRT-based estimations are needed in CAT applications (Thompson & Weiss, 2011). There are different IRT models, such as one parameter logistic model (1PLM), two parameters logistic model (2PLM), three parameters logistic model (3PLM), and these kinds of models are used when item responses are evaluated dichotomously (correct/incorrect or yes/no). Some other models are used when item responses are evaluated polytomously (Brown, 2018), and some of these models are the Partial Credit Model, Generalized Partial Credit Model, Graded Response Model, and Nominal Response Model (Doğan & Aybek, 2021). For dichotomously scored item responses in CAT applications, 3PLM is the most accepted model (Green et al., 1984; Lord,1980; Weiss, 1983; as cited in. Hambleton et al., 1991). Even individuals encounter different items in CAT

* PhD., TED University, Faculty of Education, Ankara-Türkiye, merve.kursad@tedu.edu.tr, ORCID ID: 0000-0002-6591-0705

To cite this article:
Şahin-Kürşad, M. (2023). The effects of different item selection methods on test information and test efficiency in computer adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 33-46. https://doi.org/10.21031/epod.1140757

Received: 5.07.2022
Accepted: 25.12.2022

applications, IRT models provide standards to estimate different individuals' abilities (Hambleton et al., 1991).

Computer adaptive tests are implemented on individuals using various algorithms. These algorithms are implemented in three stages that are the start, continuation, and termination stages. CAT applications start with an item selected from the item pool. The individuals' ability is estimated depending on whether they answer correctly, and another item is selected from the item pool. Ability is estimated after each time the individual receives another item. This process continues until a certain level of accuracy is reached in terms of ability estimation if the variable-length test termination rule is selected, and the process stops when a defined number of items are reached if the fixed-length test termination rule is selected. Then the test is concluded. (Eggen, 2004).

A well-defined item pool with high-quality items is necessary for CAT applications to obtain reliable and valid results, as the results are reliable and valid to the extent of the quality of items. In relation to the item pool, the item selection methods are one of the important factors that affect ability estimation, test information and test efficiency (Han, 2009). Item selection methods have been developed to make the use of the item pool more efficient (Boyd, 2003). These methods influence the start, continuation, and termination of the application of CAT because these methods determine which items the individuals should encounter according to their ability level. Aside from the CAT process, another factor that increases the importance of item selection methods is test security. These methods affect test security by providing balance in item pool usage (Sulak, 2013). There are various item selection methods used in CAT applications. Some of these are Interval Information Criterion (IIC), Efficiency Balanced Information (EBI), Matching-b Value (MbV), Kullback-Leibler Information (KLI), Maximum Fisher Information (MFI), Likelihood Weighted Information (LWI), Random Selection (RS), a-Stratification and Gradual Maximum Information Ratio (GMIR) (Han, 2012). These methods are better or stronger than each other in their various aspects, and there are some studies about the comparison of performances of these methods.

Wen et al. (2000) simulation study on the efficiency of MFI and a-stratification with 5,000 individuals and an item pool of 400 items found that the a-stratification method leads to better results compared to the MFI method in terms of efficient use of the item pool. Costa et al. (2009) compared the performance of MFI, KLI and Maximum Expected Information (MEI) methods in the CAT application. The researchers created an item pool of 246 items from the University of Brasilia's English Language Test and performed a simulation study. They used the bias and mean squared error factors while comparing these methods. It was found that all three methods produced lower bias and standard error values as the test length increased.

In Han's (2009) study examining the item selection methods' influence on item use frequency, test information, item pool index, bias and errors of Ɵ ability estimation in CAT environment, randomly selected MFI, fade-away MFI, GMIR and fade-away GMIR methods were used. According to the simulation study with 250 individuals and 500 items, it was found that although the MFI method was frequently preferred, the GMIR method produced lower standard errors in terms of the variables specified in CAT applications. Han (2010) compared the a-stratification, IIC, LWI, KLI and GMIR methods in the CAT application. The study tried to identify the item selection method which provided the most efficient use of the item pool and the best performance balance in evaluating the performance of the methods. The study was performed as a simulation study and used 500 multiple-choice items from the Graduate Management Admission Test (GMAT) and the number of individuals was simulated as 80,000, 40,000 and 20,000. According to the results of the study, the MFI, KLI and GMIR methods performed better in situations where the test length is relatively shorter.

Sulak (2013) compared the simulative performance of the MFI, a- stratification, LWI, GMIR and KLI methods. The study was simulated with 2,000 individuals and 250 items and examined the item use frequency of the item selection methods. It was concluded at the end of the study that with regards to item pool use, all item selection methods chose items with high discrimination with greater incidence and therefore were not well balanced in terms of item pool use.

Boztunç Öztürk and Doğan (2015) used a- stratification, MFI and GMIR methods in their study, where they compared different item exposure control methods. They made comparisons regarding estimation precision and test security in the study. According to one of the results of the study, MFI and GMIR methods yielded similar results in terms of item pool use, while the a-stratification method balanced item pool use even in situations where the item pool is not controlled and reduced the test overlap ratio.

There were few studies that examined the effect of item selection methods on TIF, and these studies didn't directly examine the TIF, but standard errors, which are the function of TIF, were calculated (Costa et al., 2009; Deng et al., 2010; Han, 2009; Han, 2010; Sulak, 2013; Sulak & Kelecioğlu, 2019). Also, CAT conditions (start, continuation, and termination stages), sample sizes, and item pool sizes in these studies were different from each other and this study. As mentioned above, IRT models are useful for CAT applications. One of its advantages for CAT is the estimation of individual standard error (Sulak, 2013). The standard error is used to calculate TIF, and TIF resembles how well the test is in estimating the ability and how well the test differentiates the individuals at related ability levels. In the TIF estimation process, item information is summed at related ability levels, and so, selected item information by item selection methods can affect the TIF (Baker, 1986; Hambleton et al., 1991). Therefore, it is important to investigate the effect of item selection methods on TIF.

Besides TIF, the item pool is important for CAT applications. Developing a qualitative item pool is a difficult process, and it is essential to use these items in CAT. For test security, items in the item pool should be used equally (Davis, 2002). One of the factors which affect this process is the item selection methods. These methods are important because selecting the item which maximizes the TIF about individuals taking the test is of critical importance to get effective ability estimations (Sulak, 2013). To evaluate the performance of item selection methods in the aspect of test efficiency, generally average test length (ATL), average exposure rate of items (AERI), scaled chi-square value (X^2), underexposed item rate (UIR) and overexposed item rate (OIR) values are evaluated articles in the accessible literature (Boztunç Öztürk & Doğan, 2015; Boztunç Öztürk, 2014; Han, 2009; Lee & Dodd, 2012; Moyer et al., 2012). ATL represents the number of items implemented to each individual; AERI represents the ratio of the total exposure rate of an item to the number of individuals, scaled chi-square value (X^2) represents the difference between the observed and expected item use frequency. This value gives information about how effectively the item pool is used (Chang & Ying, 1999). Underexposed Item Rate (UIR) represents the average exposure rate of items at lower than 0.02; that is, some items are rarely used, and UIR gives the rates of rarely used items. Overexposed Item Rate (OIR) represents the average exposure rate of items at higher than 0.20; that is, some items are used so frequently in CAT, and OIR gives the rate of frequently used items (Eggen, 2001). Investigation of the effect of item selection methods on these factors is vital because item selection methods can affect the usage of items that is balanced in the item pool with which item is applied next, and so it can affect the test security (Sulak, 2013).

A general evaluation of the studies mentioned above, and the literature indicates that the performance of item selection methods varies under different conditions in CAT applications. Generally, item use frequency, test length, bias and root mean square error (RMSE) values were taken into account when comparing item selection methods. Aside from this, the methods examined were the frequently used a-stratification, MFI and KLI methods (Balta & Uçar, 2022; Costa et al., 2009; Sulak, 2013; Wen et al., 2000). The purpose of this study is to compare the effects of the IIC, EBI, matching -b value, LWI and RS methods besides the frequently used a-stratification, MFI and KLI methods on test efficiency and TIF. These item selection methods were chosen because, firstly, a- stratification, MFI and KLI methods were frequently used methods in CAT applications and these methods' effect on TIF and test efficiency is important. Also, IIC, EBI, matching -b value, LWI and RS methods were selected because their item selection algorithms are different from each other and their effect on TIF and test efficiency were not investigated directly in the available literature. Besides these, research in related literature has shown that item selection methods have some advantages and disadvantages according to each other, but in these researches, two item selection methods were compared (Balta & Uçar, 2022; Wen et al., 2001; Yi & Chang, 2003) or five item selection methods were compared via only average test length in the aspect of test efficiency (Sulak & Kelecioğlu, 2019). Also, these studies' CAT application rules (start, continuation and termination stages) were different from each other, and they evaluated test efficiency

generally via only average test length. In this study, especially effect of item selection methods on test efficiency factors is investigated more detail. Accordingly, the research questions of this study are:

1. How do TIF values change according to IIC, EBI, MbV, KLI, MFI, LWI and RS item selection methods?
2. How do the average test length (ATL), average exposure rate of items (AERI), scaled chi-square value (X^2), underexposed item rate (UIR) and overexposed item rate (OIR) change according to IIC, EBI, MbV, KLI, MFI, LWI and RS item selection methods?

Method

This section provides information about the research model and data production and analysis.

Research Model

This research is a simulation study carried out with simulated data. Simulation studies are conducted under conditions where there is no real data or when situations that are more complex than real life are examined (Ranganathan & Foster, 2003). Due to the application difficulty of CAT applications and the requirement of a large item pool and large sample size, most studies on CAT are carried out using simulated data (Babcock & Albano, 2012; Deng, Ansley & Chang, 2010; Han, 2010; Sulak, 2013). As this study is also based on IRT-based CAT applications and requires a large item pool, a large sample size and is difficult to apply, it was conducted as a simulation study.

Data Production and Analysis

Production and analysis of research data were carried out through the SimulCAT (Version 1.2) program developed by Han (2012). Large-scale tests and features of CAT applications were taken into consideration in data production. Ability parameters were produced first, followed by item parameters. Constants and independent variables taken into account in simulated data production are shown in Table 1.

Table 1

Constants and Independent Variables Taken Into Account in Simulated Data Production in CAT

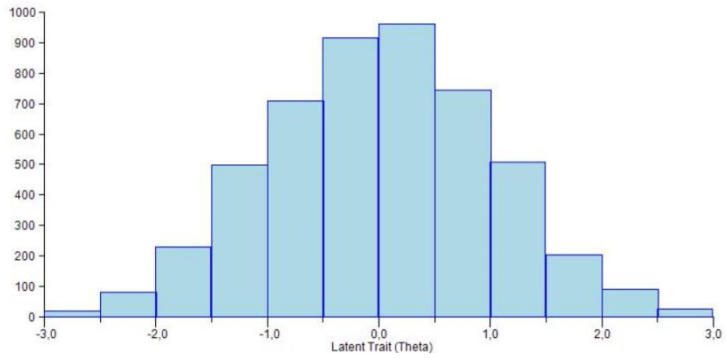
Constant Conditions	Independent Variables
1. Sample size	**1.** Item selection method
2. Distribution of Ability Parameters	**a.** Interval Information Criterion (IIC)
3. Item pool size	**b.** Efficiency Balanced Information (EBI)
4. IRT model and Distribution of Item Parameters	**c.** Matching –b Value
5. Ability estimation method	**d.** Kullback-Leibler Information (KLI)
6. Starting rule	**e.** Maximum Fisher Information (MFI)
7. Item use frequency control method	**f.** Likelihood Weighted Information (MFI)
8. Stopping rule	**g.** Random Selection

Constants of the Study Taken into Account in Simulated Data Production

Sample Size: It has been claimed that a sample size of at least 1,000 individuals is necessary to accurately perform parameter estimations in IRT-based CAT applications (Rudner & Guo, 2011; Stahl & Muckle, 2007). Besides this, in the study conducted by Şahin (2012), where different IRT models (1PLM, 2PLM, 3PLM) were compared, it was stated that low standard error values were obtained in a sample size of 5,000. Therefore, this study is based on a situation where the sample size is 5,000.

Distribution of Ability Parameters: Ability parameters were produced through a normal distribution with a mean of zero and a standard deviation of one. The reason for choosing normal distribution is that the Expected A Posteriori (EAP) method was selected as the ability estimation method. When the EAP method is used, ability parameters need to be chosen from a universe whose distribution is known, or they need to display normal distribution (Gershon, 2005). The distribution of ability parameters in a sample of 5,000 is given in Figure 1.

Figure 1
Distribution of Ability Parameters in a Sample of 5,000



Item Pool Size: Although it has been stated that an item pool of at least 100 items is necessary to make accurate ability estimations in CAT applications (Urry, 1977), an item pool with such a low number of items is not enough (Sulak, 2013). Stocking (1992) states that the item pool should be at least more than six to ten times the length of the test. In this study, the size of the item pool was set at 500. Because as stated by Risk (2010), an item pool of 500 items is taken to be the ideal item pool size for item pool certificate programs.

IRT model and Distribution of Item Parameters: As the item pool produced for this study consists of dichotomously scored (1/0 answer pattern) items, 3PLM was selected as the IRT model. According to 3PLM, the probability of answering an item correctly is calculated as follows (Birnbaum, 1968);

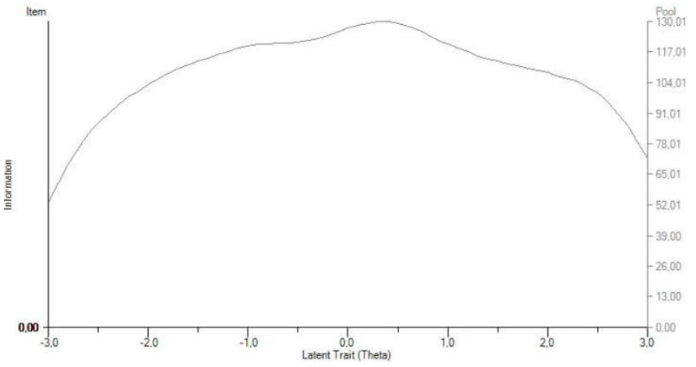
$$P\ (x_j = 1 \mid \Theta_k, a_j, b_j, c_j) = cj + \frac{1-cj}{1+ e^{-Daj\,(\theta k - bj)}} \qquad (1)$$

In the formula 1, xj represents the answer to item j (one if true, zero if false), aj represents the slope related to item j, i.e. discrimination, bj represents the threshold parameter for item j, i.e. difficulty, cj represents the low asymptote of item j, i.e. the chance parameter; while the D value represents the scaling constant and has the 1.7 constant value in normal ogive models.

Meanwhile, the distribution of item parameters can vary according to the purpose of the test. According to Boyd (2003), items in the item pool should contain various items ranging from easy to difficult for achievement tests. The ideal distribution of item parameters should consist of a uniform distribution. Therefore, as item production was carried out taking into account achievement tests and the 3PLM was utilized within the scope of this study, the -a parameter was produced from a uniform distribution with the value of between 0.50-2.00; the -b parameter was produced from a uniform distribution with the value of between -3.00-3.00; and the -c parameter was produced from a uniform distribution with the value of between 0.05-0.20. The TIF for the item pool of 500 items according to the specified item parameter values is shown in Figure 2.

Figure 2

The Test Information Function for the Item Pool of 500 Items



Ability Estimation Method: There are many ability estimation methods that are used in CAT applications, such as Weighted Likelihood Estimation (Warm, 1989), Minimum Chi Quadrant (Zwinderman & van den Wollenberg, 1990) etc., two of the most frequently used ones are Maximum Likelihood Estimation (MLE) (Baker, 1992) and Bayesian estimation methods. Bayesian estimation methods come in two types, the Expected A Posteriori (EAP) (Bock & Aitkin, 1981; Bock & Mislevy, 1982) and the Maximum A Posteriori (MAP) (Samejima, 1969). In most of the studies where these methods are compared to each other, it was found that the EAP ability estimation method produced a lower standard error and lower bias values compared to MLE and MAP methods and performed better (Eroğlu, 2013; Keller, 2000; Kezer, 2013; Kingsbury & Zara, 1989). Aside from these studies, Sulak's (2013) study, it was found that when the EAP method is utilized, different item selection methods' standard values produced lower standard errors compared to other ability estimation methods. Therefore, the EAP ability estimation method was used in the current study.

Starting rule: The starting rule covers the selection process of the first items to be selected at the beginning of the test. There are various starting rules in this context. One of them is assigning a value that corresponds to the average ability level. In CAT applications, this value is generally accepted as zero, meaning they start with average items that address zero ability levels. However, when this rule is employed, all individuals receive the same first item, and this can be a problem in the aspect of test security or item exposure. Therefore, random assignment of items can be needed. To eliminate this situation, the starting rule can choose from items with a value of between -0.5 and 0.5. With this method, it is assumed that no information is available on examinees and only several possible items, still not too much, can be assigned to individuals to estimate their initial ability (Thompson & Weiss, 2011). Therefore, this study employs a starting rule that uses items with a value of between -0.5 and 0.5, which corresponds to the average ability level.

Item use frequency control method: Item use frequency methods are used to prevent the repeated use of a certain item. The goal here is to protect the security of the item pool (Davis & Dodd, 2005). There are three items that use frequency control methods named the randomesque method, the Sympson-Hetter method and the fade-away method. In Randomesque method, items are selected from a group of items from the most informative items at the current ability level. However, this method is less successful for dichotomously scored items (Han, 2011; Kingsbury & Zara, 1989; Lee & Dodd, 2012). Sympson-Hetter method gives the conditional probability of an item which will be applied (Boztunç Öztürk, 2014; Han, 2011). Despite the Sympson-Hetter method is well in item pool usage, it is stated that it has a laborious process (Boztunç Öztürk & Doğan, 2015). In the fade-away method, less used and frequently used items' usage are balanced that is frequently used items are used less (Han, 2011). The fade-away method was used in this study as it provides more accurate results compared to other methods in terms of reducing the skewness and test conflict with regard to the item pool usage (Boztunç Öztürk & Doğan, 2015; Davis & Dodd, 2005; Han, 2012).

Stopping rule: The stopping rule refers to the test's termination criterion. There are two stopping rules that are fixed-length and variable-length stopping rules. Studies indicate that the variable-length stopping rule provides a better measurement quality compared to the fixed-length stopping rule and terminates the test more economically (Babcock & Weiss, 2012; Eroğlu, 2013). Used with the stopping rule based on the number of items, it is stated that the distribution approaches normal when 13 or more items are used (Blais & Raiche, 2010). Aside from this, it is stated that better results are obtained in terms of measurement accuracy when the standard error is equal to or lower than 0.40 in the [-3.00; +3.00] ability level range (Babcock & Weiss, 2012; Eroğlu, 2013). Therefore, this study takes into account a situation where the minimum number of items is 15, and the standard error is lower than 0.40 for the test termination rule.

Independent Variables in Simulated Data Production

Item selection methods were taken into account as an independent variables in simulated data production. The explanation of some of these methods, according to Han (2012), is given below:

Maximum Fisher Information (MFI): The MFI method is the most frequently used method and tries to find the item that maximizes the $I_i[\widehat{\Theta}_{m-1}]$ value for the m-1 item applied until that instance after each answer given to the items. It selects items through the "local information" around the relevant Θ (Weiss, 1982).

Interval Information Criterion (IIC): The IIC method was developed as an alternative to the MFI. In this method, developed by Veerkamp and Berger (1997), the information function is centered along the confidence interval of the interim Θ estimation. The information functions are averaged along the confidence interval.

Kullback-Leibler Information (KL, Kullback-Leibler Information (Global Information)): The KL method, developed by Chang and Ying (1999), is a function of two ability variables (Θ and Θ_0) and expresses the change of an item between two Θ levels and uses "global information" in item selection.

Likelihood Weighted Information Criterion (LWI): In this method, the information function is collected along the Θ scale and weighted with the probability function.

While IIC, KL and LWI methods perform evaluations along a Θ range using the item information functions, the MFI method selects items based on its evaluation according to the Θ value at a certain point.

Randomization: In this method, items are chosen randomly. As items are selected randomly, this method does not have adaptive testing features (Han, 2012). Despite of its non-adaptive feature, this method was used in research (Choi & Swartz, 2009; Eggen, 2012). For instance, Choi and Swartz (2009) compare the item selection methods as; methods which select the next item randomly and methods which select the next item based on MFI. Eggen (2012) stated that when it was used as an item selection method, efficiency loss was huge, and it affected root mean square error value negatively, especially when selecting harder items. It was stated that random selection or selection of harder items might be motivating in learning systems, which negatively affects ability estimations (Eggen, 2012). Also, the random selection method showed more bias, especially when theta values were away from the mean (Choi & Swartz, 2009). In these studies, the effect of random selection on TIF and test efficiency were not studied. Therefore, in this study, it is used to see its effect on TIF and test efficiency, also.

a-Stratification: The aim of this method is to prevent the selection of high-discrimination items. In this method, developed by Chang and Ying (1999), the items in the item pool are stratified based on their item discrimination, and the item with the -b parameter value close to the interim Θ value is selected from these strata starting with the lowest a parameter value.

Best Matching b-Value (MbV): This method is a special application of the a-stratification method. This method uses only one item stratum and the -b parameter closest to the interim Θ value is selected regardless of the -a parameter and -c parameter.

Gradual Maximum Information Ratio (GMIR): The GMIR method developed by Han (2009) is a method based on expected item efficiency. This value is defined as the effectuation level of information belonging to the item that is closest to the interim Θ.

Efficiency Balanced Information (EBI): In the EBI method developed by Han (2012), unlike the GMIR method, item information is evaluated throughout the Θ interval instead of a certain Θ point. The width of the Θ interval is determined with the standard error of estimation.

The methods included in this study, among those listed above, are the IIC, EBI, MbV, KLI, MFI, LWI and RS methods. GMIR and a-stratification methods were excluded from the study as they failed to initiate the CAT process under specified stopping rules and item pool characteristics.

A total of seven different item selection methods were compared in simulated data production. Twenty-five replications were performed for each item selection method, and 175 analyses were conducted.

The ATL, AERI, X^2, UIR, OIR were calculated while determining test efficiency and also TIF values were calculated for comparing the seven-item selection method during the data analysis process, and these can be named as dependent variables of this study. Calculations were performed through standard errors in the TIF. The formulas regarding the data analysis process are given in Table 2.

Table 2
Criteria for Evaluating Test Efficiency and Test Information Function

	Criteria	Description	Formula
Test Information Function	**TIF**	Equal to the information function of the items in the test taken by individuals in the CAT application	$\text{Sem}(\hat{\Theta})=\frac{1}{\sqrt{I(\theta)}}$
Test Efficiency	**Average Test Length (ATL)**	The number of items implemented for each individual	$\sum_{i=1}^{n}\frac{Ki}{n}$
	Average exposure rate of items (AERI)	The ratio of the total exposure rate of an item to the number of individuals	$\sum_{i=1}^{n}\frac{mk*100}{n}$
	Scaled chi-square (X^2)	Difference between the observed and expected item use frequency	$\sum_{I=1}^{N}\frac{(Koj-\underline{K}oj)^2}{\underline{K}oj}$
	Underexposed item rate (UIR < 0.02)	Average exposure rate of items at lower than 0.02	$\sum_{i=1}^{n}\frac{mk*100}{n}<0.02$
	Overexposed item rate (OIR > 0.20)	Average exposure rate of items at higher than 0.20	$\sum_{i=1}^{n}\frac{mk*100}{n}>0.20$

n: total number of individuals, *Ki*: i. the number of items implemented to each individual, *mk*: the number of times item k is applied to all individuals, *Koj*: the number of applications of the item j/number of individuals, $\bar{K}oj$: Test length/the size of item pool, I(Θ): item information function

Table 2 shows that when examining the effect of item selection methods on the TIF, the TIF is calculated based on the standard error calculated for each individual because there is an inverse relationship

between the TIF and the standard error of measurement (van der Linden, 1998). As the indicator of test efficiency value, ATL, AERI, X^2, UIR (UIR < 0.02), and OIR (OIR > 0.20) values were calculated.

Results

The results of the comparison of seven different item selection methods' TIF and test efficiency values are given below. The TIF values obtained according to item selection methods are given in Figure 3.

Figure 3
The TIF Values Obtained According to Item Selection Methods

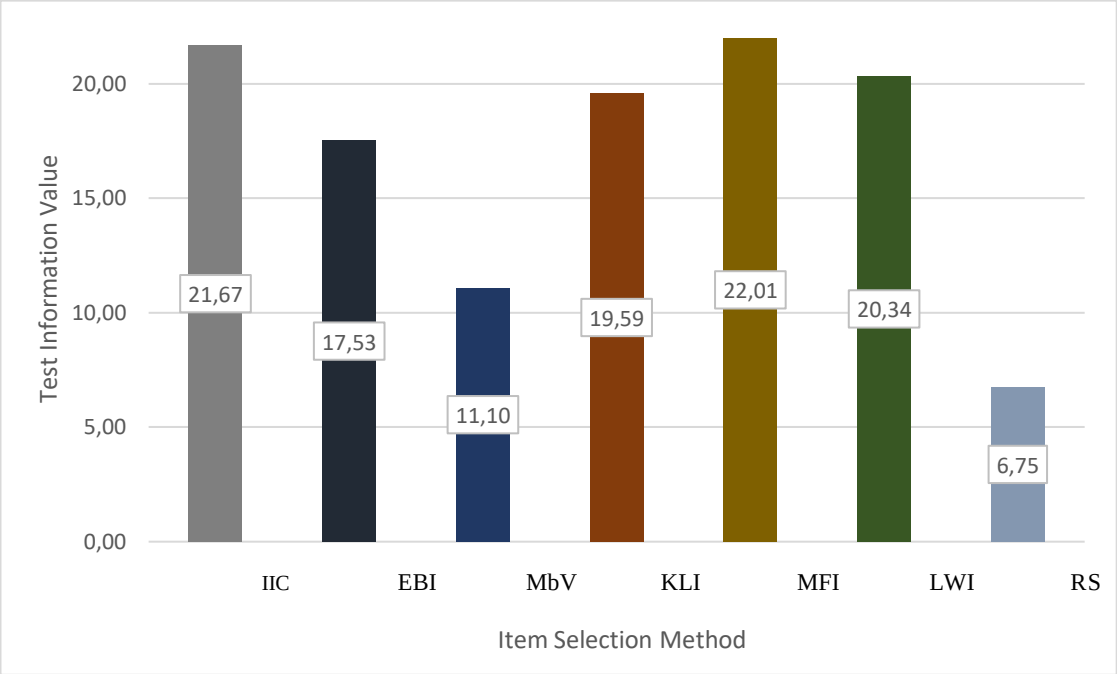


Figure 3 shows that the method which gives the highest TIF compared to other item selection methods is the MFI method (22.01) (%95 CI [22.017-21.993]. The IIC method (21.67) (%95 CI [21.661-21.682]) follows the MFI method. The method which produces the lowest test information value among the item selection methods is the RS method (6.75) (%95 CI [6.749-6.754]) which is the expected. The method which produces the lowest test information value after RS is the MbV method (11.10) (%95 CI [11.085-11.104]). The other methods which follow the MbV are EBI (17.53) (%95 CI [17.509-17.543]), KLI (19.59) (%95 CI [19.574-19.606]) and LWI (20.34) (%95 CI [20.324-20.351]) respectively. Veerkamp and Berger (1997) stated in their study that they compared the performance of the IIC, LWI and MFI methods and that the performances of these methods are close to each other as in this study.

Since the items are selected randomly in the RS method, which therefore has no adaptive test features (Han, 2012), the RS method produces the lowest test information value. The key to successful CAT implementation is choosing the best item selection method and item exposure method that will allow obtaining the best TIF (Han, 2009). It can be said that the MFI approach is used frequently in research because it is a simple, straightforward, and effective method (Han, 2009), and it was also one of the efficient methods for obtaining high test information values in this study.

The ATL, AERI, X^2, UIR and OIR values, which are indicators of test efficiency obtained according to item selection methods, are given in Table 3.

Table 3
Values for the Test Efficiency Obtained According to Different Item Selection Methods

Item Selection Method	ATL	AERI	X^2	UIR	OIR
IIC	15.00	0.03	14.94	0.05	0.81
EBI	15.00	0.03	14.94	0.03	0.73
MbV	15.00	0.03	14.94	0.00	0.51
KLI	15.00	0.03	14.94	0.06	0.83
MFI	15.00	0.03	14.94	0.05	0.79
LWI	15.00	0.03	14.94	0.05	0.80
RS	28.00	0.06	27.89	0.00	0.00

According to test efficiency values in Table 3, the ATL value is 15.00 for all item selection methods except for RS. The highest ATL value belongs to the RS method. The RS method does not produce an efficient test in terms of average test length, while other methods perform similarly. AERI represents the ratio of the total exposure rate of an item to the number of individuals. Examining the AERI values, it can be as with the ATL value, all methods perform similarly except for the RS method. As the average number of item applications was relatively higher than the others, RS (0.06) has the lowest performance.

The X^2 value represents the difference between observed and expected item use frequency. The higher this number is, the lower the test efficiency performance of the method. In other words, if this value is high, it means that some items are used repeatedly (Cheng et al., 2015). Methods other than the RS method performed close to each other, while the RS (27.89) method displayed the lowest performance as the X^2 value was higher compared to other methods. In other words, items in the item pool are used more repeatedly in the RS method compared to other item selection methods. Tested by ATL, AERI and X^2 values, it was seen that all methods except for the RS method displayed similar results. This might be related to the use of the fade-away method as the item use frequency method within the scope of this study. As Boztunç Öztürk (2014) has also stated, item use frequency ratios are close to each other since various items are used in the implementation of the fade-away method. This is supported by X^2 analysis results. Because in terms of test security, in the case the fade-away method is used as the item use frequency method, results are closer to the ideal compared to the situations where item use frequency is not controlled for (Boztunç Öztürk, 2014).

The UIR represents what percentage of items in an item pool were seen lower than 0.02 per cent. In this context, it is observed that no item (%0.00) is used in low frequency in the RS and MbV methods. Meanwhile, it is seen that the highest low exposure ratio is observed in the KLI method. The OIR represents what percentage of items in an item pool were seen higher than 0.20 per cent. In this context, the method in which the item exposure rate is highest is the KLI method, with 0.83 per cent. Following KLI is the IIC method (0.81 per cent). In the RS method (0.00 per cent), no item had a high exposure rate. In this context, it can be said that the KLI method performs in an unbalanced way in terms of underexposed and overexposed item ratios.

When low and high exposure rates are examined together, it was seen that the RS method used neither low nor high exposure items. Although the item re-use rate increases as the average test length increases in the RSI method, this increase is balanced. Among other methods with similar ATL, AERI and X^2 values, although it was seen that the KLI and IIC methods had higher low and high item exposure rates, this percentage is low. This finding indicates that those items in the pool that were used too high or too low exposure are used more frequently compared to other methods, and the item pool usage is more unbalanced, albeit very slightly (Cheng et al., 2015). This finding is also supported by the study

conducted by Sulak (2013). In the relevant study, it was stated that the lowest and highest number of item usage was obtained from the KLI method, with EAP as the ability estimation method. It is thought that a similar finding was obtained in this study as the ability estimation was performed using the EAP method.

Discussion and Conclusion

The effect of various item selection methods on TIF and test efficiency in the CAT application was examined in this study. TIF is an indicator of the amount of information provided by the test and is inversely proportional to the standard error of estimation. In other words, the information provided by the test decreases as the standard error of estimation increases (van der Linden, 1998). In this context, RS is the method that gives the highest standard error, while the MFI is the method with the lowest standard error. In the study where Han (2010) compared the performances of a-stratification, IIC, LWI, KLI and GMIR methods, it was stated that the MFI, KLI and GMIR methods displayed lower standard error values compared to other methods. It was stated that these methods give higher test information, particularly in cases where the test length is shorter. Aside from this, the MFI method being an efficient method in terms of increasing the TIF is supported by the study conducted by Han (2009). It is stated that theoretically, MFI-based methods provide maximum test information with low standard error value due to choosing items that maximize the information function (Han, 2009; 2018).

In terms of ATL, AERI and X^2 values that are the indicators of test efficiency, all methods except for the RS method performed similarly. The RS method displayed a lower performance compared to other methods in terms of test efficiency. The reason for this finding might be due to the RS method not having an adaptive test feature, as items in this method are chosen randomly without any criteria (Han, 2012). In Choi and Swartz's (2009) study comparing six different item selection methods, the RS method had the lowest correlation between the observed and expected ability values. Another situation related to the use of an item pool is that when the KLI, MFI and LWI methods which show similar performances are examined, items with higher discrimination were selected more in implementation. A similar finding can be found in a study conducted by Sulak (2013). Although it is stated that one of the ways to prevent the use of items with high discrimination is the a-stratification method (Sulak, 2013, Yi & Chang, 2003), this study did not take into account the a-stratification method. A general evaluation of the performance of methods shows that the methods other than RS have no advantage over one another both in terms of the ATL, AERI and X^2 values.

Evaluating the low and high exposure percentages of the items, it is seen that these values are lower than one and close to each other. In the study where Han (2009) compared different item selection methods, it was stated that in cases where the fade-away item use frequency is employed, the item exposure rate was more under control. The reason for low and high item exposure percentages being very low and close to each other in this study might be due to the use of the fade-away item use frequency method. In terms of item pool use frequency, MFI, a- stratification, the fact that LWI and KLI methods perform similarly is supported by the work of Sulak (2013). It was seen that the item's high exposure percentage was high in the KLI and IIC methods. Although this percentage is not very high, it poses a threat to test security because it means some items are shown to individuals at a far higher rate due to the unbalanced use of the item pool. This is a factor that threatens test security. For this reason, although it is a low ratio under the conditions taken into account in this study, the use of KLI and IIC methods should be approached with caution due to concerns about item pool security.

The purpose of the CAT applications is to measure the subject with a high validity-reliability and high-test information at the ability level of each individual. In this context, CAT is a tool to produce high test information and efficient tests suitable for the ability level of each individual. It may be suggested that, in terms of the item selection method, the MFI method performs better compared to other methods, and the RS method has the worst performance and should not be selected because the RS method also has not an adaptive feature. The MbV method followed the RS method in the aspect of worse performance in TIF. Since methods other than the RS and MbV display similar performances in terms of TIF and test efficiency, any one of them can be chosen. However, in order to obtain better results in terms of TIF, it

is suggested that implementers use the MFI method. For further studies, it may be suggested that as this study worked with a constant sample and item pool size, they should examine the effects of these methods in a smaller or larger sample and item pool sizes. Aside from these, a constant start rule, termination rule, ability estimation method and item use frequency was employed in the CAT implementation process in this study. The effects of item selection methods on ability estimations and test efficiency can be examined by modifying the specified CAT conditions.

Declarations

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

Ethical Approval: The data used in this study were generated by simulation. Therefore, ethical approval is not required.

References

Babcock, B. & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*(7), 565- 580. https://doi.org/10.1177/0146621612455090

Babcock, B. & Weiss, D.J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CAT's provide efficient and effective measurement? *International Association for Computerized Adaptive Testing, 1*, 1-18. http://dx.doi.org/10.7333%2Fjcat.v1i1.16

Baker, F. (1986). The basics of item response theory. *Journal of Educational Measurement, 23*(3), 267-270.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. Marcel Dekker.

Balta, E., & Uçar, A. (2022). Investigation of measurement precision and test length in computerized adaptive testing under different conditions, *E-International Journal of Educational Research, 13*(1), 51-68. https://doi.org/10.19160/e-ijer.1023098

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chaps. 17–20). AddisonWesley.

Blais, J. & Raiche, G. (2010). Features of the sampling distribution of the ability estimate in Computerized Adaptive Testing according to two stopping rules. *Journal of applied measurement, 11*(4), 424-31.

Bock, R. D. & Aitkin, M. (1981).Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://link.springer.com/article/10.1007/BF02293801

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431– 444. https://doi.org/10.1177/014662168200600405

Boyd, M. A. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* [Unpublished Doctoral Thesis]. The University of Texas.

Boztunç Öztürk, N. (2014). *Bireyselleştirilmiş bilgisayarlı test uygulamalarinda madde kullanım sıklığı kontrol yöntemlerinin incelenmesi [Investigatıon of item exposure control methods in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Boztunç Öztürk, N. & Doğan, N. (2015). Investigating item exposure control methods in computerized adaptive testing. *Educational Sciences: Theory and Practice, 15*(1), 85-98. https://doi.org/10.12738/estp.2015.1.2593

Brown, A. (2018). Item response theory approaches to test scoring and evaluating the score accuracy. In Irwing, P., Booth, T. & Hughes, D. (Eds.), *The Wiley Handbook of Psychometric Testing*. John Wiley & Sons.

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

Choi, S. W. & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419–440. https://doi.org/10.1177/0146621608327801

Cheng, Y., Patton, J.M. & Shao, C. (2015). a-Stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement, 75*(2), 260-283. https://doi.org/10.1177/0013164414530719

Costa, D., Karino, C., Moura, F. & Andrade, D. (2009, June). *A comparision of three methods of item selection for computerized adaptive testing* [Paper Presentation] The meeting of 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Çıkrıkçı-Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test [New horizons in psychometrics: Individualized test in computer environment]. *Türk Psikoloji Bülteni, 5*(13), 31-36.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items.* [Unpublished Doctoral Dissertation], The University of Texas.

Davis, L. L., & Dodd, B. G. (2005). *Strategies for controlling item exposure in computerized adaptive testing with partial credit model.* Pearson Educational Measurement Research Report 05-01.

Deng, H., Ansley, T. & Chang, H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*(2), 202-226. https://www.jstor.org/stable/20778948

Doğan, C.D. & Aybek, E.C. (2021*). R-Shiny ile psikometri ve istatistik uygulamaları [Psychometric and statistical applications with R-Shiny]*. Pegem Akademi.

Eggen, T.J.H.M. (2001). Overexposure and underexposure of items in computerized adaptive testing. Measurement and Research Department Reports, 2001-1. Citogroep

Eggen, T.H.J.M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Print Partners Ipskamp B.V., Citogroup Arnhem.

Eggen, T.H.J.M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. *Psychometrics in Practice at RCEC*, 11.

Eroğlu, M.G. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması [Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*(1), 109–127.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational measurement*, *21*(4), 347-360. https://www.jstor.org/stable/1434586

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

Han, K.T. (2009). *Gradual maximum information ratio approach to item selection in computerized adaptive testing*. Council Research Reports, Graduate Management Admission.

Han, K.T. (2010). *Comparison of Non-Fisher Information Item Selection Criteria in Fixed Length Computerized Adaptive Testing* [Paper Presentation] The Annual Meeting of the National Council on Measurement in Education, Denver.

Han, K. T. (2011). *User's Manual: SimulCAT.* Graduate Management Admission Council.

Han, K.T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement*, *36*(1), 64-66.

Han, K. (2018). Components of item selection algorithm in computerized adaptive testing. *J Educ Eval Health Prof, 15*(7). https://doi.org/10.3352/jeehp.2018.15.7

Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması [Comparison of adaptive (individualized) test application and traditional paper-pencil test application in ability estimation]* [Unpublished Doctoral Dissertation]. Hacettepe University

Keller, A.L. (2000). *Ability estimation procedures in computerized adaptive testing*. Technical Report, American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams.

Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [Comparison of computerized adaptive testing strategies]* [Unpublished Doctoral Dissertation]. Ankara University.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement, 72*(1), 159-175. https://doi.org/10.1177/0013164411411296

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

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement, 72*(4). https://doi.org/10.1177/0013164411431838

Ranganathan, K. & Foster, I. (2003). Simulation studies of computation and data scheduling algorithms for data grids. *Journal of Grid Computing, 1*, 53-62. https://doi.org/10.1023/A:1024035627870

Risk, N.M. (2010). *The impact of item parameter drift in computer adaptive testing* (CAT) [Unpublished doctoral dissertation]. University of Illinois.

Rudner, L.M. & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Graduate Management Council (GMAC)*, 11(01), 6-10.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement, 34*(4). https://doi.org/10.1002/j.23338504.1968.tb00153.x

Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması [Comparision of item selection methods in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(3), 315-326. https://doi.org/10.21031/epod.530528

Stahl, J. A. & Muckle, T. (2007, April). *Investigating displacement in the Winsteps Rasch calibration application* [Paper Presentation] The Annual Meeting of the American Educational Research Association, Chicago, IL.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm*. Research Report 93-2, Educational Testing Service.

Şahin, A. (2012). *Madde tepki kuramında test uzunluğu ve örneklem büyüklüğünün model veri uyumu, madde parametreleri ve standart hata değerlerine etkisinin incelenmesi [An investigation on the effects of test length and sample size in item response theory on model-data fit, item parameters and standard error values]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation.*, 1-9. https://doi.org/10.7275/wqzt-9427

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196. https://www.jstor.org/stable/1434014

van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201–216. https://doi.org/10.1007/BF02294775

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Some New Item Selection Criteria for Adaptive Testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226. https://doi.org/10.3102/10769986022002203

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20. http://dx.doi.org/10.1111/j.1745-3992.1993.tb00519.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (1982). Latent Trait Theory and Adaptive Testing. In David J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 5-7). Academic Press.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–-375. https://www.jstor.org/stable/1434587

Wen, H., Chang, H. & Hau, K. (2000). *Adaption of a-stratified Method in Variable Length Computerized Adaptive Testing*. American Educational Research Association Annual Meeting, Seattle.

Yi, Q. & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*, 359–378. https://doi.org/10.1348/000711003770480084

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14*(1), 73-81. https://doi.org/10.1177/014662169001400107

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 47-61

Developing A Computerized Adaptive Test Form of the Occupational Field Interest Inventory

Volkan ALKAN* Kaan Zülfikar DENİZ**

Abstract

In this research, the aim was to apply the Occupational Field Interest Inventory (OFII), which was developed in paper-pencil format, as a Computerized Adaptive Test (CAT). For this purpose, the paper and pencil form of the OFII was applied to 1425 high school students and post-hoc simulations were carried out with the obtained data. According to results obtained from the simulations, it was decided that the most ideal criteria for the CAT application were GPCM as the IRT model, .40 standard error value as the test termination rule, and MFI as the item selection method. The OFII ended with an average of 59 items, and the correlations between scores obtained from the paper-pencil form and thetas (θ) estimated by simulation ranged between .91-.97. According to post-hoc simulation results, the CAT application was applied to 150 students. It was observed that the correlations between the scores of students from the online application of the paper-pencil form and θ levels estimated by the CAT form varied between .73 and .91.

Keywords: Computerized Adaptive Test, Item Response Theory, Occupational Field Interest Inventory, Occupational Interest

Introduction

Having an occupation is an important factor for people to maintain their lives to a certain standard by obtaining the necessary income to do so, which can also play an important role in determining an individual's social prestige as well as their achievement of happiness (Altın, 2020). While choosing an occupation suitable for oneself, individuals often make their choice based on comparing their personal knowledge (i.e., lifestyle, interest, skills, values, etc.) with the available occupations as well as the conditions of those occupations (Akar, 2012). According to Yoo (2016), the factors which affect an individual's career choice can be listed as occupational interest, talent, personality, value, socioeconomic status, and gender. Among these factors, one of the variables that most affects an individual's career choice is occupational interest.

Occupational interest is initially determined by an individual's liking for people who do a specific job. For example, occupational interests are determined by assuming that someone who loves teachers will in effect have an interest in the teaching occupation. Later, this method was abandoned, and occupational interests were then determined according to the individuals' enjoyment of behaviors belonging to various occupations (Deniz, 2009). Today, occupational interest is mostly determined by asking individuals about their level of interest in a range of work activities through inventories.

When the measurement tools used to measure occupational interest were examined, it could be seen that most of them were developed based on Classical Test Theory (CTT). CTT has been widely used in measurement applications such as test development, application, and evaluation since the early 1900s (Hambleton et al., 1991). In CTT, the sum of scores that an individual obtains from items of the measurement tool are defined as the degree of possessing the feature to be measured. Although there are some exceptions, a low score that the individual obtains from the measurement tool generally indicates

* PhD., Ankara University, Faculty of Education, Ankara-Türkiye, volkanalkan114@gmail.com, ORCID ID: 0000-000-8264-8190
** Prof. Dr., Ankara University, Faculty of Education, Ankara-Türkiye, kzdeniz@ankara.edu.tr, ORCID ID: 0000-0003-0920-538X

To cite this article:
Alkan, V., & Deniz, K. Z. (2023). Developing a computerized adaptive test form of the Occupational Field Interest Inventory. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 47-61. https://doi.org/10.21031/epod.1153713

Received: 3.08.2022
Accepted: 15.02.2023

that the level of possessing the desired feature being measured is low, while a high score indicates that the level of having the desired feature being measured is high. CTT applications are mostly concerned with test-level information such as reliability. In addition, it should be noted that although CTT allows for obtaining item-level information such as item discrimination index and average of item scores, CTT does have important limitations. The limitations of CTT are that item statistics are dependent on the group, test scores obtained by individuals are dependent on the test items, the inability to distinguish individuals being different from the average in terms of ability level, and measurement error is considered the same for all individuals while measurement error is actually different for each individual (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Meyer, 2010). As a result, over time the limitations of CTT have been discussed and many have sought a new model to eliminate these limitations. Thus, the model developed by taking these limitations into account is the Item Response Theory (IRT).

IRT is an item-based theory based on the psychological measurement studies of Binet, Simon and Terman in 1916. The first studies regarding IRT were made by Thorndike, Thurstone, Horst and Symonds in the 1920s, and in the subsequent years, Lord, Novic and Lawley continued studies regarding IRT as well as significantly contributed to the theory's development (Ostini & Nering, 2010). IRT, especially as a result of developments in computer technology, has frequently been used in the measurement of various characteristics within the field of psychology and education, and has also been developed as an alternative for addressing limitations arising from the structure of CTT (Harvey & Hammer, 1999).

IRT has many application areas, and in particular, one of these application areas is the Computerized Adaptive Test (CAT), which was developed using IRT. CAT is a computer-based application in which each individual test-taker does not answer the same items, but instead only the items appropriate to their skill-feature levels as measured within the test (Kezer & Koç, 2014).

The development of CAT models suitable for both two-category and multi-category items have brought to the forefront the idea of developing CAT-forms for the measurement of items normally applied in a paper-pencil format. Thus, valid and reliable measurement tools, which can be difficult to implement in terms of application time, are applied in a shorter amount of time through the use of CAT applications due to fewer items being needed than in the paper-pencil form (Özbaşı & Demirtaşlı, 2015). In addition, with the help of CAT, it is possible to perform more reliable measurements in a shorter amount of time by not querying individuals using items that are well above or well below their ability level (Şahin & Özbaşı, 2017). Furthermore, some measurement tools may need updating according to the technological and social developments experienced, which may ultimately reduce the usefulness of the scales by causing an increase in the number of items used within the measurement tools. In this respect, the adaptation of valid and reliable measurement tools originally applied in a paper-pencil format to a CAT format also facilitates updating studies to be carried out regarding these scales.

The Occupational Field Interest Inventory (OFII), which is the subject of this research, is one of the inventories for which the CAT application had yet to be developed. The OFII is an interest inventory which includes 14 subscales, consisting of 156 items, and the paper-pencil application of this inventory takes approximately 15-20 minutes. When the CAT application for this inventory was developed, it was clear that the usefulness of the inventory would increase by shortening the application time as well as new sub-scales belonging to different occupations could be added. The primary purpose of the current research was to further develop the CAT form by determining the most appropriate IRT model, test termination rule, and item selection method for the OFII, which was developed to assist students in their career choices as part of student occupational guidance services. GPCM Generalized Partial Credit Model (GPCM) and Graded Response Model (GRM) were preferred as the IRT model, .30, .40, .50 as the standard error value and Maximum Fisher Information (MFI), Maximum Expected Information (MEI), Minimum Expected Posterior Variance (MEPV) and Maximum Expected Posterior Weighted Information (MEPWI) as the item selection method. These are preferred because the platform used in the research allows working with these options, and these options are generally preferred in scales developed in accordance with multi-category models (Aybek & Çıkrıkçı, 2018; Boyd et al., 2010;

Özbaşı, 2014; Şimşek 2017; Van der Linden, 1998). In line with the determined general purpose, answers were sought to the following questions.

1. When different IRT models (GPCM and GRM) and three different standard error values (.30, .40, and .50) and different item selection methods (MFI, MEI, MEPV, and MEPWI) are used as test termination rules:
   a) How many items are used on average in the OFII-CAT (OFII-C) simulations?
   b) How do the standard error values to be obtained from the OFII-C simulations change?
   c) How does the direction and level of the relationship change between the θ levels obtained from the OFII-C simulation and the paper-pencil application of the OFII (OFII-PP)?
2. When the OFII-C application is created using the test termination rule, item selection method, and IRT model determined according to the findings obtained from the OFII-C simulation results:
   a) How does the frequency of item use change in OFII-C and OFII-C simulation applications?
   b) How do the average number of items and test times change in the OFII-C and OFII paper-pencil form online application (OFII-PPOA)?
   c) Is there a significant relationship between the θ levels obtained by the students from the OFII-C application and the scores they received from the OFII PPOA?

Method

Information regarding the research model, research group, data collection tool, data collection process, and data analysis are presented in this section.

Research Groups

There were two different research groups which took part in this study. The first research group was the one with whom the OFII-PP application was carried out, and the data based on the post-hoc simulation application was obtained. In this group, there were 1425 students from the 10th, 11th or 12th grade studying at different types of high schools located in Mersin, Turkey during the 2018-2019 academic year. The second research group is the group in which both OFFI-PPOA and OFFI-C applications were carried out. In this group, there are 150 students studying in the 10th, 11th and 12th grades of different types of high schools in Mersin. Data were collected from this research group in April 2020.

Data Collection Tools

The OFII-PP developed by Deniz (2009) was used in the first data collection phase of this research. In the second phase, OFII-PPOA (which is the computerized version of OFII) and OFII-C, which was developed as a result of post-hoc simulations, were used. The OFII is an inventory aimed at assisting individuals in selecting an occupation. To develop OFII, first, by making use of the university's educational programs and literature, 14 occupational fields (i.e., computer, law, health, psychology, mathematics, literature, visual arts, foreign language, political sciences, science, communication, education, agriculture, and engineering) were reviewed, and 25 items each focused on measuring the interest of students were written. The items were rated on a 5-point Likert scale: (1) I find little interest, and (5) I find it very interesting (Deniz, 2009). To analyze the validity and reliability of the inventory, it was applied to 1373 students studying at 10 high schools in Ankara, Turkey. Thus, to determine the validity of the inventory, exploratory factor analysis was applied on 1373 students and confirmatory factor analysis was applied to a data group of 216 randomly selected students from the original group of 1373. To determine the reliability of the inventory, Cronbach alpha internal consistency coefficient was calculated with the data obtained from two separate groups of 673 and 700 people, which were determined from the original group of 1373. In addition, the inventory was reapplied to 109 students

selected from among the group of 1373 for whom the inventory was applied, and the test-retest reliability coefficient was calculated (Deniz, 2009).

To determine the content validity of OFII, expert opinion on the content validity of the sub-scales of OFII was obtained from 88 academics who work in various occupational fields at a variety of universities and have earned at least a doctorate degree within their field of expertise. Following the content validity, exploratory factor analysis was applied to determine the construct validity of OFII. As a result of confirmatory factor analysis performed using 14 subscales obtained from the exploratory factor analysis, 11 or 12 items with high factor load values belonging to the subscale were determined and a final version of the inventory consisting of 156 items was obtained. Thus, as a result of the confirmatory factor analysis, it was determined that the goodness of fit indexes of the subscales (CFI, GFI, NNFI, and AGFI) were above .90, except for the AGFI value (.87) of the science subscale. Similarly, in terms of RMSEA values, RMSEA values below .08 were obtained for all factors, except for the science subscale (.087). In addition, correlation coefficients were calculated to determine the relationships between subscales and values ranging from -.43 to .50 were obtained (median: -.07). The fact that the majority of the correlation coefficients obtained had negative values indicated that the subscales sufficiently diverged from each other. Another application carried out to determine the construct validity of the inventory was to calculate the correlation between the scores that students directly gave to their occupation names, varying between 1 and 9, and the scores they got from the relevant sub-scale of the inventory. As a result of these calculations, it was seen that there were significant positive correlations between .49 - .80 (Deniz, 2009).

To estimate the reliability of the inventory, Cronbach's alpha reliability coefficient was calculated for each subscale. Cronbach's alpha internal consistency coefficients for the subscales of OFII were found to range from between .79 and .95 (Median: .88). In addition, the test-retest reliability of the inventory was determined by re-administering the inventory to the same participants eight weeks later. After these applications, it was determined that the test-retest reliability coefficients of the subscales of OFII varied between .75 and .95 (Deniz, 2009).

Data Collection Process

In the scope of this research, an individualized form of the OFII was developed within a computer environment. The data obtained from the first research group were collected from seven high schools accessible to the researcher. Before the data were collected, permission was obtained from the Turkish Ministry of National Education (MoNE) and data were collected with the help of psychological counselors working within the schools. The data collection process was carried out in the classroom environment and each student was provided a booklet containing 156 items belonging to the OFII along with an answer sheet, and they were asked to only answer using the answer sheet. Participation in the research was strictly voluntary and students were informed that if they participated in the research, their individual results would be shared with them at a later date. After the obtained data were analyzed, the occupational field interest profiles of each student were sent to the guidance counselors within their schools and the results shared with the individual students. As a result, it was observed that the students carefully examined their occupational field interest profiles as well as shared and discussed the results with their friends.

As part of the first application, the item parameters of the OFII according to GRM and GPCM were obtained using data obtained from the OFII-PP form. Then, based on the item parameters obtained from the application, a post-hoc simulation was carried out for the CAT form via the Firestar (Choi & Swartz, 2009) software. Thus, the CAT simulation was carried out separately for each subscale, and correlation coefficients were calculated to determine the relationship between the average number of items in each subscale, the mean standard error values, θ levels estimated for all items, and θ levels estimated as a result of the simulation. According to the results, the most suitable IRT model to be used in the OFII-C application was determined as the item selection method and test termination rule. In the second phase of the research, the OFII-C application was developed on the Concerto platform and applied to 150 students online at www.meslekialanilgienvanteri.com. Similarly, it was applied to the same students

using the Google Survey Application for the OFII-PPOA. To prevent rank effect, a 15-day period wait period was carried out between the applications and 75 students who had taken the OFII-PPOA in the first application received the OFII-C within the second application.

Analysis of Data

In the analysis of the current research data, IBM SPSS Statistics 20.00, LISREL 8.51, R, Firestar and PARSCALE software were used.

Examining Assumptions

First, the one-dimensionality assumption, which is a basic assumption of the IRT, was examined through confirmatory factor analysis due to the factor structure of OFII being predetermined. Confirmatory factor analysis was performed with LISREL 8.51 software separately for each dimension and the assumptions of the analysis were checked prior to the factor analysis. Goodness-of-fit indices of the confirmatory factor analysis which were performed separately for each subscale are presented in Table 1.

Table 1

Confirmatory Factor Analysis Fit Indices Applied for the One-Dimensional Assumption

Subscale	χ^2	sd	p	χ^2/sd	RMSEA	SRMR	AGFI	NFI
Computer	58.35	44	.00	1.32	.043	.039	.93	.97
Law	74.62	44	.00	1.70	.056	.051	.91	.95
Health	62.42	44	.00	1.42	.044	.042	.92	.96
Psychology	71.88	44	.00	1.63	.053	.044	.90	.95
Math	73.39	44	.00	1.67	.055	.048	.91	.93
Literature	95.46	44	.00	2.17	.072	.065	.88	.92
Visual arts	97.63	44	.00	2.22	.073	.067	.88	.92
Foreign language	89.52	54	.00	1.66	.057	.054	.90	.94
Political science	106.48	54	.00	1.97	.065	.061	.89	.92
Science	111.04	44	.00	2.52	.082	.079	.86	.90
Communication	63.17	44	.00	1.44	.045	.040	.91	.96
Education	59.82	44	.00	1.36	.044	.040	.91	.94
Agriculture	84.25	44	.00	1.92	.067	.063	.89	.93
Engineering	74.33	44	.00	1.69	.056	.052	.92	.95

In Table 1, the goodness of fit indices obtained for each subscale of the OFII were evaluated according to the criteria determined by Schermelleh-Engel et al. (2003). As a result of the confirmatory factor analysis performed for each subscale of the OFII, χ^2/sd values was found below 3 for all subscales, indicating that the data fit perfectly with the model. In addition, it was seen that the RMSEA value for only the science subscale did not show a good fit. Although the RMSEA value obtained for the science subscale was above 0.80, the 0.90 confidence interval of the RMSEA value for the science subscale indicated that the RMSEA value of 0.082 was acceptable. Thus, according to the results obtained, it can be stated that each subscale of the OFII provided an assumption of unidimensionality.

After testing the unidimensionality assumption, the invariance of the item parameters were tested. For this purpose, two different data groups consisting of 500 people were created randomly from the data set of 1425 people. For each data group created, first of all, the item parameters were calculated using the PARSCALE software, then the relationship between the item parameters calculated for both groups were determined by calculating the Spearman Rank Differences Correlation Coefficient due to the scarcity of items in the subscales. In the next step, the items in each subscale were divided into two groups, and the relationship between the students' θ values estimated according to the items in both groups were determined using the Pearson Product Moments Correlation Coefficient.

The findings regarding the invariance of the θ estimations and item parameters for each subscale of the OFII are presented in Table 2 for both GPCM and GRM. Since the θ estimations and a parameter gave

close values for the GRM and GPCM models, only the position parameter was calculated according to the different IRT models.

Table 2

Findings on the Invariance of θ Estimates and Item Parameters

	r_a	r_L		r_θ
Subscales		GPCM	GRM	
Computer	.69	.82	.82	.43
Law	.82	.94	.94	.56
Health	.83	.94	.93	.57
Psychology	.91	.96	.95	.49
Math	.81	.91	.92	.46
Literature	.79	.92	.92	.61
Visual arts	.84	.94	.95	.67
Foreign language	.73	.88	.89	.52
Political science	.84	.95	.94	.57
Science	.59	.82	.81	.46
Communication	.76	.89	.89	.41
Education	.68	.86	.86	.65
Agriculture	.84	.92	.93	.55
Engineering	.72	.87	.89	.48

All correlation coefficients provided in Table 2 were found to be significant ($p < .05$). Accordingly, it can be stated that the item parameters and θ estimations showed the invariance feature.

Data Analysis for Post-hoc Simulation

After the item parameters were determined, the appropriate syntax was created for the simulation to be carried out in the R software using Firestar (Choi, 2009). While performing the simulations, GRM and GPCM were used as the IRT model. In the first item selection, θ = 0.00 was determined and MEPV, MEI, MEPWI, MFI were used as the item selection method. Standard error values of 0.30, 0.40, and 0.50 were preferred, provided that at least three items were used as the test termination rule. While the range [-3,3] was determined as the θ interval, the θ increment was determined as 0.10. The BS (EAP) was preferred as the θ estimation model. The mean distribution was determined as 0.00 and the standard deviation was 1.00 as the a priori distribution as well as the posterior distribution was preferred as the standard calculation method. While item use control was not carried out, the scaling value was determined as D = 1.7. As in this research, while deciding on the specified simulation conditions, studies were used in which the CAT form of an affective measurement tool was developed and successful results were obtained (Aybek & Çıkrıkçı, 2018; Şimşek, 2017). In addition, the limitations of the Concerto application, in which the OFII-C application is carried out, were considered.

According to the simulation result, for each subscale, the average number of items the application ended with was determined as well as the average standard error and correlation coefficients between the θ values obtained as a result of the simulation and the θ values obtained from the whole test were obtained. In addition, according to the results obtained, it was decided which test termination rule, item selection method, and IRT model were most suitable for the OFII-C.

Data Analysis for OFII-C Application

The Pearson Product-Moment Correlation Coefficient was used to calculate the correlation between the θ levels estimated from the OFII-C application and the scores obtained from the OFII-PPOA application. Furthermore, frequency of item use was determined, and frequency analysis was performed for the items used. After calculating the correlation coefficients and the frequency of item use, the OFII-PPOA scores and the OFII-C estimations were provided in the same graph as a way of comparing the occupational field interest profiles obtained from the OFII-PPOA and OFII-C applications. For this purpose, the raw scores obtained by the students from the OFII-PPOA form were converted into standard z scores.

Thus, to determine whether the OFII-PPOA and OFII-C profiles matched, the relationship between the θ levels obtained by the students in the second research group of 150 people from the OFII-C application and the scores they obtained from the OFII-PPOA application were determined by calculating the Pearson correlation coefficient.

Results

The first sub-objective of this research was to determine the mean number of items used, the mean standard error values obtained, and the mean standard error values obtained in the post-hoc simulations performed using different IRT models (GPCM and GRM) along with different test termination rules (0.30-0.40-0.50 standard error). The correlations between the θ estimates are presented in Table 3.

Table 3

GRM and GPCM Values for OFII .30, .40 and .50 Standard Error Test Termination Rules

		k		SEM		r	
	Subscales	GPCM	GRM	GPCM	GRM	GPCM	GRM
SEM= 0.30	Computer	10.12	11.00	.32	.41	.99	1.00
	Law	9.52	11.00	.31	.32	.91	1.00
	Health	10.00	11.00	.31	.31	.98	1.00
	Psychology	11.00	11.00	.33	.39	1.00	1.00
	Math	8.43	11.00	.31	.31	.81	1.00
	Literature	9.41	11.00	.31	.47	.88	1.00
	Visual arts	11.00	11.00	.33	.35	1.00	1.00
	Foreign Language	10.34	12.00	.32	.36	.99	1.00
	Political science	7.86	11.45	.30	.38	.66	.99
	Science	10.05	11.00	.32	.39	.98	1.00
	Communication	6.28	10.57	.30	.40	.56	.99
	Education	6.71	10.52	.30	.45	.59	.99
	Agriculture	10.20	11.00	.32	.33	.99	1.00
	Engineering	8.27	11.00	.30	.35	.79	1.00
SEM= 0.40	Computer	4.80	9.15	.38	.43	.94	.98
	Law	4.12	8.27	.36	.40	.93	.97
	Health	4.20	8.37	.37	.41	.93	.97
	Psychology	5.25	8.12	.38	.41	.96	.99
	Math	3.53	7.48	.36	.41	.91	.96
	Literature	3.81	8.06	.36	.40	.92	.96
	Visual arts	5.00	9.93	.38	.41	.96	.98
	Foreign Language	5.53	9.74	.38	.41	.95	.98
	Political science	3.29	6.87	.35	.39	.91	.95
	Science	4.30	8.50	.37	.39	.94	.98
	Communication	3.18	6.31	.35	.38	.90	.95
	Education	3.22	6.51	.35	.39	.91	.94
	Agriculture	5.11	9.29	.38	.42	.94	.98
	Engineering	3.41	7.11	.36	.40	.91	.96
SEM= 0.50	Computer	4.54	8.70	.41	.46	.90	.95
	Law	3.93	7.16	.39	.44	.89	.93
	Health	3.97	7.75	.40	.45	.89	.93
	Psychology	5.17	9.95	.43	.48	.92	.96
	Math	3.17	6.76	.39	.44	.89	.93
	Literature	3.29	6.96	.39	.44	.89	.93
	Visual arts	4.81	9.81	.43	.48	.91	.95
	Foreign Language	4.32	9.56	.43	.48	.91	.95
	Political science	3.15	6.28	.40	.46	.88	.93
	Science	4.00	8.20	.40	.46	.90	.94
	Communication	3.01	6.05	.37	.43	.88	.92
	Education	3.00	6.16	.38	.43	.88	.92
	Agriculture	4.57	9.09	.42	.46	.90	.94
	Engineering	3.15	6.58	.38	.43	.88	.93

Thus, according to the post-hoc simulation results, when a standard error value of .30 was preferred as the test termination rule, it was seen that the mean standard error was above this value in all subscales. Whereas when a standard error value of .40 was used as the test termination rule, it was determined that an average standard error of over .40 was obtained for the subscales of GRM except in the subscales of political sciences, science, communication, and education. However, it was determined that GPCM had a standard error of less than .40 in each subscale as well as this occurred by using approximately 4.2 items. When the standard error value of .50 was preferred as the test termination rule, an average standard error value of less than .50 was obtained for all subscales in both the GRM and GPCM.

As a result of the calculation made using the data obtained from the OFII-PP application, it was determined that the Cronbach's alpha internal consistency coefficients of the subscales ranged between .81 and .94. When the .40 standard error was selected as the test termination rule, the average of the reliability coefficients of the reliability scales became .84.

According to the results of the post-hoc simulation research, a standard error of .40 was found to be more appropriate as a test termination rule than other rules, and as a result, the decision was made to use the .40 standard error as test termination rule in the OFII-C application. In addition, according to the simulation studies carried out, it was seen that GPCM achieved similar results with fewer items than GRM. Also, it was determined that GPCM used approximately 62% fewer items than the 156 items in the original form as well as provided feature estimation with an error of less than .40. Due to all of these reasons, the decision was made to use GPCM as the IRT model in the OFII-C application.

Thus, with a .40 standard error value as the test termination rule, the GPCM and MFI, MEI, MEPV, and MEPWI item selection methods as the IRT model, the correlations between all test-simulation θ estimations obtained as well as the standard error values and average number of items used are presented in Table 4.

Table 4

Findings According to Item Selection Methods According to 0.40 Standard Error Value as Test Termination Rule

Subscale	MFI			MEI			MEPV			MEPWI		
	k	SEM	r	k	SEM	r	k	SEM	r	k	SEM	r
Computer	4.80	.38	.94	4.82	.38	.94	4.83	.38	.94	4.84	.38	.94
Law	4.12	.36	.93	4.12	.36	.93	4.13	.36	.93	4.13	.36	.93
Health	4.20	.37	.93	4.20	.37	.93	4.21	.37	.94	4.22	.37	.94
Psychology	5.25	.38	.96	5.27	.37	.94	5.26	.37	.94	5.27	.37	.94
Math	3.53	.36	.91	3.54	.36	.92	3.53	.36	.92	3.54	.36	.92
Literature	3.81	.36	.92	3.81	.36	.92	3.80	.36	.92	3.80	.36	.92
Visual arts	5.00	.38	.96	5.02	.38	.96	5.02	.38	.97	5.02	.38	.97
Foreign Language	5.53	.38	.95	5.53	.38	.94	5.52	.38	.94	5.53	.39	.95
Political science	3.29	.35	.91	3.28	.35	.93	3.27	.35	.94	3.28	.35	.94
Science	4.30	.37	.94	4.31	.37	.94	4.32	.37	.94	4.34	.37	.94
Communication	3.18	.35	.90	3.18	.35	.90	3.19	.35	.90	3.19	.35	.90
Education	3.22	.35	.91	3.22	.35	.92	3.21	.35	.92	3.21	.35	.92
Agriculture	5.11	.38	.94	5.12	.36	.94	5.11	.36	.94	5.13	.37	.94
Engineering	3.41	.36	.91	3.43	.36	.92	3.42	.37	.92	3.43	.37	.92

When Table 4 is examined, it can be seen that different item selection methods did not cause a significant change in the correlation coefficients between the θ levels (all-θ) estimated using the entirety of the items and the θ levels (sim-θ) estimated by simulation as well as the standard error values or average number of items applied. Therefore, in the OFII-C application, the MFI method was preferred as the appropriate item selection method.

In addition, when GPCM was preferred as the IRT model, a .40 standard error as the test termination rule, and MFI preferred as the item selection method, it was seen that the OFII-C simulation ended with an average of 59 items. Considering that the OFII-PP consisted of 156 items, it can be stated that

approximately 62% less items were used with the OFII-C simulation. Furthermore, the correlation coefficients between the sim-θ and all-θ were determined to be at values between .91 and .97.

For the second sub-purpose of this research, the students' data for the OFII-C application were taken from a database of the website created by the researcher for the OFII-C application. Using the data obtained from that database, the frequency of use of each item was determined. Similarly, the frequency of use of the items in the post-hoc simulation application were also determined, and the frequency of use of the items according to both applications are compared in Table 5.

Table 5

Subscales of OFII Item Use Frequencies for OFII-C and OFII-C Simulation Applications

			Comp.	Law	Heal.	Psy.	Math	Lit.	Vis.	Fore.	Pol.	Sci.	Com.	Edu.	Agri.	Eng.
Item 1	Live	%	58.5	48.1	52	62.5	35.1	69.5	100	0	0	100	14	14.4	39.5	11.7
	Sim	%	62.6	44.4	43.6	53.2	25.6	60	81.6	5.7	2.5	82.5	11.3	11.2	32.7	9.2
Item 2	Live	%	12.3	56.2	74.5	54.7	43.4	55.9	12.4	69.5	0	31.7	17.1	58.3	14.3	16.2
	Sim	%	25.2	60	60.2	61.5	40.7	55.1	9.5	60.2	8.5	26.4	14.3	47.9	11.1	15.78
Item 3	Live	%	52.3	20.3	14.3	39.2	25.6	17.2	15.6	75.6	0	25.0	25.1	60.1	33.7	35.6
	Sim	%	47.8	30.3	22.5	35.5	22.2	20.5	11.4	71.2	6.1	22.0	18.5	55.2	27.5	30.1
Item 4	Live	%	24.6	100	21.3	100	100	0	33.8	0	100	17.5	100	100	42.4	100
	Sim	%	17.8	86.5	20.3	84.7	89.5	4.6	31.2	4.2	76.5	19.3	92.3	85.2	38.9	89.1
Item 5	Live	%	15.6	15.1	19.0	28.3	0	42.0	46.7	17.3	48.7	23.4	0	25.6	0	25.2
	Sim	%	20.4	20.2	14.0	22.6	7.33	50.6	40.1	15.3	39.5	27.6	6.2	30.8	8.5	28.9
Item 6	Live	%	100	91.2	5.29	14.5	0	100	0	43.6	57.8	52.2	13.4	12.5	100	14.4
	Sim	%	87.2	85.1	7.25	18.9	4.5	92.9	6.6	37.4	65.1	45.8	17.7	17.5	89.5	19.5
Item 7	Live	%	45.7	18.4	12.2	12.3	26.7	0	39.0	26.7	0	24.5	0	24.7	9.25	0
	Sim	%	51.6	15.0	15.39	15.1	28.5	3.4	34.6	23.4	11.4	32.7	10.1	33.0	14.3	5.9
Item 8	Live	%	72.0	32.7	23.6	14.6	13.5	12.4	7.6	0	0	20.2	60.5	11.5	39.5	49.1
	Sim	%	65.2	29.5	20.9	20.0	16.1	15.3	13.9	6.9	5.9	25.5	54.6	9.2	32.2	53.0
Item 9	Live	%	0	0	11.2	30.3	0	0	14.7	0	61.2	13.3	32.5	0	44.1	39.8
	Sim	%	5.08	7.58	8.83	27.6	5.5	7.8	12.3	2.5	63.5	10.4	25.1	2.5	35.4	34.1
Item 10	Live	%	33.6	27.5	65.5	0	49.5	74.2	51.6	43.7	0	17.1	45.1	45.8	66.1	43.8
	Sim	%	43.5	31.0	46.1	6.54	45.7	70.0	57.3	35.6	8.2	20.5	33.4	38.2	70.3	38.5
Item 11	Live	%	33.5	22.4	100	51.3	61.0	65.9	67.9	0	60.2	52.5	54.5	41.4	12.5	20.0
	Sim	%	37.0	19.3	92.2	45.0	67.3	54.1	59.4	6.2	65.8	42.4	47.2	43.8	9.3	23.3
Item 12	Live	%								31.5	54.3					
	Sim	%								25.8	50.2					

Live: OFII-C Sim: OFII-C Simulation

When Table 5 is examined, it can be seen that all the items were used in the post-hoc simulation, but that some items were not used in the OFII-C application. As a result, items not used in the OFII-C application were the ninth item for the computer factor; ninth item for the law factor; tenth item for the psychology factor; fifth, sixth, and ninth items for the mathematics factor; fourth, seventh, and ninth items for the literature factor; sixth item for the visual arts factor; first, fourth, eighth, ninth, and eleventh items for the foreign language factor; first, second, third, seventh, eighth, and tenth items for the political sciences factor; fifth and seventh items for the communication factor; ninth item for the education factor; fifth item for the agriculture factor; and seventh item for the engineering factor. Thus, it was seen that the frequency of use of a majority of the items which were never used in the OFII-C application was below 10% within the post-hoc simulation. In addition, it can be seen that the α parameters of the items which were never used, generally belonged to items with the lowest coefficient in each factor. When the

data of the OFII-C application was examined, it could be seen that one item in each factor was directed to all the participants. Since the $\theta = 0$ was chosen as the initial value of the test, the starting material was the same for all participants and the frequency of use of these items was determined to be 100%. In other words, in the OFII-C application, it was determined that one item in each factor was directed to all the participants. At the same time, the frequency of use of these items in the OFII-C simulation was over 80%.

Thus, to compare the average number of items used in the OFII-C application and the OFII-PPOA, the average of the number of items answered by each participant for each factor was determined through the OFII-C application. The average number of items used in the OFII-C application and the OFII-PPOA is presented in Table 6.

Table 6

Number of Items Used in OFII-C and OFII-PPOA and Test Durations

	OFII-C				OFII-PPOA	
Subscale	Minimum Number of Items	Maximum Number of Items	Average Number of Items	Average Test Time (minutes)	Average Number of Items	Average Test Time (minutes)
Computer	3	5	3.22	.29	11	.92
Law	3	5	3.35	.31	11	.95
Health	3	5	3.14	.28	11	1.02
Psychology	3	7	3.43	.31	11	.91
Math	3	6	3.74	.33	11	.94
Literature	3	5	3.42	.30	11	.88
Visual arts	3	6	3.53	.34	11	.90
Foreign Language	3	7	4.14	.36	12	1.10
Political science	3	7	4.19	.37	12	1.15
Science	3	7	3.52	.31	11	.99
Communication	3	5	3.34	.30	11	.85
Education	3	5	3.72	.33	11	.96
Agriculture	3	7	3.45	.31	11	1.00
Engineering	3	6	3.63	.32	11	.97
Total			53.49	4.46	156	13.51

When Table 6 is examined, it can be seen that the least used item subscale in the OFII-C application was the health subscale, and that the most used item subscale was the political sciences subscale. In the OFII-C application, each participant responded to an average of 53.49 items. Since the OFII-PPOA consisted of 156 items, it was stated that 65.71% less items were used with the OFII-C. While the participants answered the OFII-C within an average of 4.46 minutes, the average response time for the OFII-PPC was 13.51 minutes. In this respect, it could be stated that the application time of the OFII decreased by 66.99% with the CAT application.

The Pearson correlation coefficients were calculated to determine whether there was a significant relationship between the θ levels estimated by the students within the OFII-C application and the scores they had obtained from the OFII-PPOA. Thus, the results of the correlation coefficients are presented in Table 7.

Table 7

The Correlations Between Levels of θ Estimated from OFII-C Form and Scores Obtained from OFII-PPOA

Subscale	r	p
Computer	.88	.00
Law	.83	.00
Health	.92	.00
Psychology	.82	.00
Math	.79	.00
Literature	.91	.00
Visual arts	.85	.00
Foreign Language	.78	.00
Political science	.84	.00
Science	.74	.00
Communication	.79	.00
Education	.81	.00
Agriculture	.73	.00
Engineering	.76	.00

When Table 7 is examined, it can be seen that the correlation coefficients calculated for the relationship between the OFII-C and OFII-PPOA were significant for all the subscales and that the highest correlation coefficient was 0.92 for the health subscale, while the lowest correlation coefficient was 0.73 for the agriculture subscale. The median value of the obtained correlation coefficients was found to be 0.82. As a result, it can be stated that there were highly significant relationships between the OFII-C and OFII-PPOA for all subscales.

Discussion and Conclusion

The most important difference of the CAT applications from paper-pencil applications is that the number of items directed to each person differs. This is due to the test termination rule, which is one of the basic components of CAT applications. For example, according to the test termination rule, following each item answered by a test taker, it is determined whether the test should be terminated or continue (Hambleton et al., 1991). Although there are many different options which make up the test termination rule in CAT applications, the most preferred rule is the use of the standard error value. In the current research, the standard error criterion was used as a rule for terminating the test. There is a negative relationship between the standard error and measurement precision. As the standard error increases, the measurement precision decreases. Thus, by making use of the relationship between measurement accuracy and standard error, the standard error criterion is determined to obtain the desired measurement precision (Özkan, 2014). As a result, in cases where the standard error criterion is applied as the test termination rule, when a participant answers an item, it is determined whether the calculated standard error value is less than the determined critical value. If the standard error value calculated with this method is less than the standard error value determined as the critical threshold, the CAT application is terminated. It is revealed in several past studies, (Babcock & Weiss, 2012; Eroğlu & Kelecioğlu, 2015; Gnambs & Batinic, 2011; Stochl et al., 2016), that the test termination rule is one of the most important CAT components which directly affects test length. When the literature was examined, it was recognized that the standard error criteria of 0.30, 0.40, and 0.50 were most widely used in studies (Aybek & Çıkrıkçı, 2018; Özbaşı & Demirtaşlı, 2015; Şimşek 2017) which investigated the adaptability of affective measurement tools for CAT. Therefore, a standard error criteria of 0.30, 0.40, and 0.50 were also used in this research. Thus, according to the results of the post-hoc simulation research, the .40 standard error was found to be more appropriate as a test termination rule than other rules, and as a result, the decision was made to use the .40 standard error as the test termination rule for the OFII-C application. In addition, according to the simulation studies carried out, it was seen that GPCM achieved similar results with fewer items than GRM. Furthermore, it was ultimately determined that GPCM used approximately 62% fewer items than the 156 items from the original form as well as gave feature

estimation with an error of less than .40. For the reasons just discussed, the determination was made to use GPCM as the IRT model for the OFII-C application.

In CAT applications, a variety of methods are used to determine which items are directed to the individual test taker following the selection of the starting material. For example, Van der Linden (1998) states that if one of the MEI, MEPV or MEPWI methods is selected in CAT applications, the θ estimation will be more reliable than other methods. It is also recommended by Boyd et al. (2010) that MFI, MEI, MEPV, and MEPWI methods be preferred as the item selection methods in CAT applications. Therefore, the MFI, MEI, MEPV, and MEPWI methods were the preferred item selection methods for this research. It has been recognized that different item selection methodologies do not cause a significant change in correlation coefficients, standard error values or the average number of items applied between θ levels estimated using all items (all-θ) and θ levels estimated through simulation (sim-θ). These findings were in line with the finding from Choi and Swartz (2009) using the CTM model, which the estimated θ level and number of items used in cases where the item pool is small do not differ according to the item selection method. Also, Veldkamp (2003) states that although different item selection methods are used, the same items are found at a rate between 85% to 100%. While Aybek & Çıkrıkçı (2018) used MEPWI, MEI, MFI, and MEPV item selection methods, and find that item selection methods do not have a significant effect on the estimated θ level, standard error values, and number of items used. In the current research, in accordance with the literature, it was determined that different item selection methods under both the GPCM and GRM models did not cause a significant change in the estimated θ level, standard error values, and the number of items used. Therefore, in the OFII-C application, the MFI method was the preferred item selection method.

When GPCM was preferred as the IRT model, .40 standard error as the test termination rule, and MFI as the preferred item selection method, it was seen that the OFII-C simulation ended with an average of 59 items. Considering that the OFII-PP consists of 156 items, it can be stated that approximately 62% less items were used with the OFII-C simulation. In addition, it was determined that the correlation coefficients between sim-θ and all-θ gained values between .91-.97. In Scullard (2007), an investigation of the adaptability of the Strong Interest Inventory, a measurement tool similar to OFII for individuals in the computer environment, reached the same findings obtained from our research. The correlation coefficient obtained in Scullard (2007) ranged from .90 to .98 between the sim-θ and all-θ estimations and the test length decreased by approximately 60%. The fact that many researchers (Betz & Turner, 2011; Chien et al., 2011; Gibbons et al., 2012; Hol et al., 2007; Smits et al., 2011) obtained similar findings to those obtained in the current research highlights the reliability of the findings.

The OFII-C was developed through the Concerto program, using GPCM as the IRT model, .40 standard error value as the test termination rule, and MFI as the item selection method, which are the most suitable criteria for OFII-C. The OFII-C and OFII-PPOA were applied to 150 high school students. While all items were used in post-hoc simulations, it was determined that some items were not used in the OFII-C application. It was also observed that the majority of the items which were never used in the OFII-C application were the items with a very low frequency of use in the post-hoc simulation. In addition, it was determined that the α parameters of the items which were never used, generally belonged to the items with the lowest coefficient in each factor.

It was determined that approximately 66% less items were used in the OFII-C application compared to the OFII-PPOA, and that the OFII-C was 67% more advantageous in terms of time. When the literature was examined, it was found that there were findings similar to the current research regarding the OFII-C application ending with fewer items as well as a shorter time period compared to the OFII-PPOA application. For example, Hol et al. (2007) adapted the measurement tool from a paper-pencil format to a CAT format and had a 62.50% decrease in the number of items. While Gibbons et al. (2012) showed a 95% decrease in the number of items for their 626-item measurement tool adapted to the CAT format. Also, when Kocalevent et al. (2009), adapted a 104-item measurement tool to the CAT format, the decrease in the number of items was 85%. Whereas Aybek and Çıkrıkçı (2018) adapted their measurement tool to the CAT format and had a 52% decrease in the number of items. In addition, Şimşek (2017) adapted the measurement tool in the CAT format and had a 50% decrease in the number of items. Thus, in a variety of other studies, it can be seen that the CAT application saves between 50-70% in the

number of items used as well as an overall decrease in test time (Betz & Turner, 2011; Bulut & Kan, 2012; Choi et al., 2010; Cömert, 2008; İşeri, 2002; Jodoin et al., 2006; Kalender, 2012; Kezer & Koç, 2014; McDonald, 2002; Öztuna, 2008; Rezaie & Golshan, 2015; Scullard, 2007; Smits et al., 2011; Weiss, 2011).

It was ultimately determined that there were highly significant relationships between OFII-C and OFII-PPOA for all subscales, and although correlation coefficients varying in the range of 0.91-0.97 were obtained between θ levels obtained from the OFII-PP and OFII-C simulation applications; the correlation coefficients varying in a range of 0.73-0.91 were obtained between the θ levels obtained from the OFII-C and OFII-PPOA applications. Thus, in this case, it can be stated that although the relationship between the OFII-C and OFII-PPOA was high, it was relatively lower than the correlation coefficients obtained from the OFII-PP and OFII-C simulation studies. In addition, when the literature was examined, it was seen that the correlation coefficients obtained in post-hoc simulation studies were higher than in the CAT studies (Achtyes et al., 2015; Aybek & Çıkrıkçı, 2018; Betz & Turner, 2011; Gibbons et al., 2012; Simms & Clark, 2005; Stochl et al., 2016).

As a result of this study, CAT form of OFI was developed successfully within the limitations of the research. The simulation phase of the research was limited to 0.30, 0.40, 0.50 standard error values, FEYB, BEYB, BEDSV, BEYSAB item selection methods and GKPM and KTM models due to the program used. At the same time, existing subscales of OFI were used while creating the CAT form and no new subscales were added. In line with these limitations, researchers can be recommended to develop CAT form of OFI by using different standard error values, item selection methods, IRT models, and adding new professional interests that have emerged in accordance with the technological developments of our age.

Declarations

Author Contribution: Volkan Alkan: conceptualization, investigation, methodology, data curation, supervision, writing - review & editing. Kaan Zülfikar Deniz: 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: The study was ethically approved by the Ministry of National Education (research number: 81576613/605.01/5603857, dated 18/03/2019). In addition, this study was found ethically appropriate with the decision of Ankara University Rectorate Ethics Committee numbered 56786525-050.04.04/49481. This study has been produced from the dissertation of Volkan Alkan that was conducted under the supervision of Prof. Dr. Kaan Zülfikar Deniz.

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

Achtyes, E. D., Halstead, S., Smart, L., Moore, T., Frank, E., Kupfer, D. J., & Gibbons, R. D. (2015). Validation of computerized adaptive testing in an outpatient nonacademic setting: The vocations trial. *Psychiatric Services*, 1-6. https://doi.org/10.1176/appi.ps.201400390

Akar, C. (2012). Factors affecting university choice: A study on students of economics and administrative sciences. *Journal of Eskişehir Osmangazi University Faculty of Economics and Administrative Sciences, 7*(1), 97-120.

Altın, M. (2020). Education, status and social mobility in Turkey. *Mecmua, 10*, 180-196. https://doi.org/10.32579/mecmua.789249

Aybek, E. C., & Çıkrıkçı, R. N. (2018). Applicability of self-assessment inventory as an individualized test in computer environment. *Türk Psikolojik Danışma ve Rehberlik Dergisi*, *8*(50), 117-141.

Babcock, B., & Weiss, D. (2012). Termination criteria in computerized adaptive tests: Do variable - length CATs provide efficient and effective measurement? *Journal of Computerized Adaptive Testing*, *1*(1), 1-18. https://doi.org/10.7333/1212-0101001

Betz, N. E., & Turner, B. M. (2011). Using Item Response Theory and Adaptive Testing in Online Career Assessment. *Journal of Career Assessment*, *19*(3), 274–286. https://doi.org/10.1177/1069072710395534

Boyd, A., Dodd, B., & Choi, S. (2010). Polytomous models in computerized adaptive testing. Nering, M., & Ostini, R. (Ed.). *Handbook of polytomous item response theory models*. (229-255). Routledge.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research*, (49), 61–80.

Chien, T.-W., Lai, W.-P., Lu, C.-W., Wang, W.-C., Chen, S.-C., Wang, H.-Y., & Su, S.-B. (2011). Web-based computer adaptive assessment of individual perceptions of job satisfaction for hospital workplace employees. *BMC Medical Research Methodology*, *11*(1), 1-8. https://doi.org/10.1186/1471-2288-11-47

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program forpolytomous item response theory models. *Applied Psychological Measurement*, *33*(8), 644–645.

Choi, S. W., & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement*, *33*(6), 419-440. https://doi.org/10.1177/0146621608327801

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), 125–136. https://doi.org/10.1007/s11136-009-9560-5

Cömert, M. (2008). *Development of computer-aided assessment and evaluation software adapted to the individual*. Unpublished Master's Thesis, Bahçeşehir University Institute of Science and Technology, İstanbul.

Deniz, K. Z. (2009). Occupational Interest Inventory (OFII) development study. Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, *6*(1), 289-310.

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

Eroğlu, M. G., & Kelecioğlu, H. (2015). Comparison of Different Termination Rules in terms of Measurement Accuracy and Test Length in Individualized Computerized Testing Applications. *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, *28*(1), 31-52.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. a, Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104-12. https://doi.org/10.1001/archgenpsychiatry.2012.14

Gnambs, T., & Batinic, B. (2011). Polytomous adaptive classification testing: Effects of item pool size, test termination criterion, and number of cutscores. Educational and Psychological Measurement, 71(6), 1006–1022. https://doi.org/10.1177/0013164410393956

Hambleton, R., & Swaminathan, R. (1985). *Fundementals of item response theory.* Sage Publications, Inc.

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

Harvey, R. J., & Hammer, A. L. (1999). Item response theory. *The Counseling Psychologist*, *27*(3), 353-383.

Hol, M. A., Vorst, H. C. ve Mellenbergh, G. J. (2007). Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison with Short Forms. *Applied Psychological Measurement*, *31*(5), 412–429. https://doi.org/10.1177/0146621606297314

İşeri, A. I. (2002). *Assessment of students' mathematics achievement through computer adaptive testing procedures*. Unpublished Doctoral Dissertation, Middle East Technical University, Ankara.

Jodoin, M. G., Zenisky, A., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, *19*(3), 203–220. https://doi.org/10.1207/s15324818ame1903_3

Kalender, İ. (2012). Computerized adaptive testing for student selection to higher education. *Yükseköğretim Dergisi*, *2*(1), 13-19.

Kezer, F, & Koç, N. (2014). Comparison of Individualized Test Strategies in Computer Environment. *Eğitim Bilimleri Araştırmaları Dergisi*, *4*(1), 145-174. https://doi.org/10.12973/jesr.2014.41.8

Kocalevent, R. D., Rose, M., Becker, J., Walter, O. B., Fliege, H., Bjorner, J. B., ... & Klapp, B. F. (2009). An evaluation of patient-reported outcomes found computerized adaptive testing was efficient in assessing stress perception. *Journal of Clinical Epidemiology*, *62*(3), 278-287.

McDonald, P. L. (2002). *Computer adaptive test for measuring personality factors using item response theory*. Unpublished Doctoral Dissertation. The University Western of Ontario, London.

Meyer, J. P. (2010). *Understanding measurement: Reliability*. Oxford University Press.

Ostini, R., & Nering, M. L. (Eds.). (2010). *Polytomous item response theory models.* Taylor and Francis Group.

Özbaşı, D., & Demirtaşlı, N. (2015). Developing the computer literacy test as an individualized test in the computer environment. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 218-237.

Özkan, Y. Ö. (2014). A comparison of estimated achievement scores obtained from student achievement assessment test utilizing classical test theory, unidimensional and multidimensional IRT. *International Journal of Human Sciences*, *11*(1), 20-44.

Öztuna, D. (2008). *Application of computer adaptive testing method in disability assessment of musculoskeletal problems*. Unpublished Doctoral Dissertation. Ankara University Institute of Health Sciences, Ankara.

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations*, *2*(5), 128–137.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

Scullard, M. G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory*. Unpublished Doctoral Dissertation. University of Minnesota, USA.

Simms, L. J., & Clark, L. A. (2005). Validation of a computerized adaptive version of the Schedule for Nonadaptive and Adaptive Personality (SNAP). *Psychological Assessment*, *17*(1), 28–43. https://doi.org/10.1037/1040-3590.17.1.28

Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. *Psychiatry Research*, *188*(1), 147–155. https://doi.org/10.1016/j.psychres.2010.12.001

Stochl, J., Böhnke, J. R., Pickett, K. E., & Croudace, T. J. (2016). An evaluation of computerized adaptive testing for general psychological distress: combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. *BMC Medical Research Methodology*, *16*(1), 58. https://doi.org/10.1186/s12874-016-0158-7

Şahin, A., & Özbaşı, Ö. (2017). Effects of Content Balancing and Item Selection Method on Ability Estimation in Computerized Adaptive Tests. *Eurasian Journal of Educational Research*, *69*, 21-36.

Şimşek, A. S. (2017). *Adaptation of skills confidence occupational interest inventory and development of computerized individualized testing*. Unpublished Doctoral Dissertation, Ankara University Institute of Educational Sciences, Ankara.

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1-23.

Van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*(2), 201-216.

Veldkamp, B. P. (2003). *Item selection in Polytomous CAT*. In Yanai H., Okada A., Shigemasu K., Kano Y., & Meulman J. J. (Eds.), New Developments in Psychometrics (pp. 207-214). Springer Verlag.

Yoo, J. H. (2016). The effect of professional development on teacher efficacy and teachers' self-analysis of their efficacy change. *Journal of Teacher Education for Sustainability*, *18*(1), 84–94. https://doi.org/10.1515/jtes-2016-0007

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 62-75

Examining The Rater Drift in The Assessment of Presentation Skills in Secondary School Context

Aslıhan ERMAN ASLANOĞLU* Mehmet ŞATA**

Abstract

The alternative assessment, including peer assessment, helps students develop metacognition among the sub-categories of assessment types. Despite the advantage of alternative assessment, reliability and validity issues are the most significant problems in alternative assessment. This study investigated the rater drift, one of the rater effects, in peer assessment. The performance of 8 oral presentations based on group work in the Science and Technology course was scored by 7th-grade students (N=28) using the rubric researchers developed. The presentations lasted for four days, with two presentations each day. While examining the time-dependent drift in rater severity in peer assessment, the many-Facet Rasch Measurement model was used. Two indexes (interaction term and standardized differences) were calculated with many-facet Rasch measurement to determine the raters who made rater drift either individually or as a group. The analysis examined the variance of scores in the following days compared to the first day's scores. Accordingly, the two methods used to determine rater drift gave similar results, and some raters at the individual level tended to be more severe or lenient over time. However, no significant rater drift at the group level showed that drifts had no specific models.

Keywords: Alternative assessment, many-facet Rasch measurement, peer assessment, validity, rater drift

Introduction

There has been an increase in the demand for qualified labor over the past century. Occupational groups desire individuals who can solve problems, think critically, analyze and present data effectively, have effective verbal and written communication skills, and evaluate themselves and their peers (Dochy, 2001). Education plays a crucial role in raising individuals; therefore, raising individuals with such characteristics can be accomplished through education systems oriented in this direction (Batmaz et al., 2022; Kaya et al., 2023). However, traditional assessment approaches applied in the learning environment are insufficient in measuring the mentioned characteristics. This new understanding necessitates establishing a connection between learning and assessment processes, which heightens the use of alternative assessment in education (Oosterhof, 2003).

Unlike traditional approaches, students are not just passive recipients of information in alternative assessment approaches. The most significant characteristic of these approaches is to make individuals develop higher-order thinking skills, such as critical and creative thinking and problem-solving skills, by actively participating in the process (Kutlu et al., 2014). Having gained importance with new approaches, performance evaluation measures how well the student uses the basic knowledge they have gained while performing complex tasks in real life. In this respect, performance assessment is unlike classical tests (multiple-choice, short answer, matching, etc.) which are concerned with the student's ability to retrieve information from memory as it is. It is based on the process of actively constructing knowledge and using it in real life. (Moore, 2009). Students should be allowed to interact with their peers and teachers during this process. Thus, it becomes possible for students to construct knowledge and share structured knowledge. Assessment and evaluation are indispensable components of learning.

* Asst. Prof. Dr., Ufuk Universty, Faculty of Education, Ankara-Türkiye, aslihanerman@yahoo.com, ORCID ID: 0000-0002-1364-7386
** Asst. Prof. Dr., Ağrı İbrahim Çeçen University, Faculty of Education, Ağrı-Türkiye, mehmetwsata@gmail.com, ORCID ID: 0000-0003-2683-4997

To cite this article:
Erman-Aslanoğlu, A., & Şata M. (2023). Examining the rater drift in the assessment of presentation skills in secondary school context. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 62-75. https://doi.org/10.21031/epod.1213969

Received: 3.12.2022
Accepted: 15.03.2023

They positively affect the learning process by improving the quality of learning and improving the learners' sense of thinking and autonomy (Orsmond et al., 2000). Alternative assessment, which includes peer assessment, has been accepted as the one that allows students to develop metacognitive skills among the subcategories of assessment types (Liu & Brantmeier, 2019).

The peer assessment emphasizes formative purposbushees to show students' performance rather than the summative purposes of assessment by enabling students to take responsibility for their learning (Azarnoosh, 2013). Studies assert that students learn better when they can benefit from the opinions of their peers (Black et al., 2003; Topping, 2017). Peer assessment can also be one of the guiding elements of group work appropriately conducted, which is essential in today's business life. In this respect, peer assessment practice carried out in group work can contribute to the success of individuals as it can increase their responsibility of individuals (Falchikov & Goldfinch, 2000; Yurdabakan & Cihanoğlu, 2009). In addition, students have versatile feedback on the quality of their work to the extent of a single instructor evaluation which is much more classical than peer assessment (Topping, 2009).

When peer assessment is used in the teaching process, the most important problem is the reliability and validity of the scores obtained from these sources (Donnon et al., 2013; Hafner & Hafner, 2003; Topping, 2003). A limited number of studies on the reliability and validity of peer assessment emphasize the importance of peer assessment in the teaching process and state that its validity and reliability are sufficient when appropriately done (Patri, 2002). In order to increase the reliability of the results obtained from peer reviews, it is of great importance to try to increase rater reliability. Validity of the results obtained from the performance measurement is required to ensure the scores' reliability (Hafner & Hafner, 2003; Jonsson & Svingby, 2007).

However, while evaluating students' performance, different factors arising from the raters can interfere with the measurement results. Rater-based factors affecting student performance are called rater effects (Farrokhi et al., 2011). Errors in rater decisions (i.e., rater effects) can influence the accuracy of assigned ratings. Although there are various errors originating from the raters in the performance assessment process, the most common rater-effective errors in the literature are rater severity and leniency and are discussed as halo effect, central tendency, range restriction, and differential rater functioning (DRF) (Myford & Wolfe, 2003). Another rater effect type that has recently attracted researchers' attention is the differential rater functioning over time (drift). Accordingly, the issue of whether the rater effect changes over time becomes significant with the increase in the use of large-scale tests, especially the implementation and scoring of which spread over time, and it has been tried to determine whether rater behaviors are affected by this situation and cause measurement errors (Congdon & McQueen, 2000; Harik et al., 2009; Myford & Wolfe, 2009).

Drift is the changes that occur in the scoring of students' performance at different times and in the rater behaviors depending on time (Wolfe et al., 2007). Various types of drift have been defined concerning the direction in which drift manifests itself. Recency drift and primacy drift are the most common (Myford & Wolfe, 2003). In Primacy drift, raters give high scores when scoring and tend to give lower scores as the rating progresses, yet Recency drift refers to the opposite. In summary, in this common type of drift, raters may display more severe or lenient behavior over time (differential severity). However, the item facet -calibration should be performed - must remain constant so that the measurement results can be compared in scoring at different times. (Leckie & Baird, 2011). Especially in cases where more than one rater implements the scoring process (i.e., peer assessment), rater calibration is not to change from person to person and from time to time (Congdon & McQueen, 2000). Such situations threaten the validity of scores by causing irrelevant variance related to students' performance (Messick, 1994).

Previous studies made several suggestions in order to reduce rater-related errors, including training raters (Hauenstein & McCusker, 2017; Knoch et al., 2007), involving more than one rater in the process (Kubiszyn & Borich, 2013), using rubrics (Andrade, 2005; Oosterhof, 2003), and adding such methods in the classroom more often (Bushell, 2006; Topping, 2003). As such, there can be less concern about the reliability of the scores. Researchers also recommended the Multi Facet Rasch Model (MFRM) to determine the reliability and validity of peer review scores (Farrokhi et al., 2011; Myford & Wolfe, 2009). MFRM removes the limitations of Classical Test Theory approaches. In evaluating students'

performance in MFRM, the factors that may affect the scores of the students are not limited to the ability levels of the individuals or the difficulty levels of the items used in the measurement process. Factors related to raters can also cause variability in student performance scores (Baird et al., 2013), which makes MFRM a suitable option for performance evaluations affected by rater behavior (Mulqueen et al., 2000).

It has been observed that the limited number of studies on drift in the literature are generally related to higher education level and second language English teaching proficiency exams. The findings of these studies, which tried to determine whether the scores given by the raters changed over time, showed different results. Some studies found rater drifts over time (Börkan, 2017; Congdon & McQueen, 2000; McLaughlin et al., 2009; Myford, 1991; Myford & Wolfe, 2009; Pinot de Moira et al., 2002). For example, Myford (1991) found rater severity due to evaluating students' drama performance for a month by referees with different experiences. Another study using MFRM determined the severity estimates of 10 raters who scored elementary school students' English writing tasks using rubrics for seven days (Condon & Mcquenn, 2000). According to the research results, it has been observed that while 9 of the raters gave more severe scores over time, one rater started to give more tolerant scores. As a result, these researchers have found rater severity for raters daily, but there was no general pattern of change. On the other hand, some researchers have not encountered scoring severity as a result of their studies (Humphris & Kaney, 2001; Leckie & Baird, 2011). For instance, Humphris and Kaney (2001) have concluded that in a 4-station exam (all stations took 5 minutes with a simulated patient) to measure the communication skills of first-year undergraduate medical students with the patient, the raters did not make rater drifts during the scoring made at different times. Similarly, in their study, Leckie and Baird (2011) concluded that the raters evaluating the scores of 14-year-old students from a large-scale English writing skill test did not make rater drifts. Some researchers, on the other hand, have obtained different results according to the structure of the exams related to rater severity. For instance, Lunz and Stahl (1990), in their study using MFRM, which took four days to score, investigated whether the raters in the oral exam, English composition exam, and clinical exams made rater drift in the scoring that lasted for four days. Research findings have shown that while rater drifts were observed in clinical and English composition exams, raters did not make any time-dependent rater drifts in the oral exam.

Purpose and Significance of the Research

Most peer assessment studies have been carried out at the higher education level, especially in foreign language teaching. The compatibility of peer scores with teacher scores is the basis for determining validity and reliability. These studies are based on the assumption that teacher scores are valid and reliable. However, teacher scores may not always be reliable and may be affected by various errors. To increase the reliability of the results obtained from peer reviews, it is of great importance to try to increase rater reliability because the validity of the results obtained from the performance measurement is possible with the reliability of the scoring (Jonsson & Svingby, 2007). Prejudices in assessment results arising from raters for different reasons threaten validity as they are sources of variance unrelated to the construct (Messick, 1994). Therefore, it is necessary to provide evidence of the validity of peer review scores. However, the evidence for the validity of peer assessment in the literature is limited (Börkan, 2017). In addition, using MFRM on rater effects provides statistical approaches to identify and correct some of these rater biases in the studies. While these approaches do not guarantee that all rater effects will be identified and deleted from test scores, they provide important information about identifying and taking action on a significant portion of rater effects.

In the present study, not only the rubrics but also the multiple raters were used in the process of measuring students' performance in peer assessment in order to satisfy the reliability of the measurement, and the MFRM approach was used.

Considering all of these, the current study attempts to explore the status of the rater effect in performance scoring when peer assessment was extended to a process. In this regard, the study, implemented with secondary school students, aims to reveal the students' behaviors in the peer-assessment process

spreading over four separate days with the help of the rubrics. In particular, the study seeks to answer the following questions:

1. Do the raters as a group demonstrate differential severity/leniency behavior throughout the scoring period?

2. Does any individual rater demonstrate a severity/leniency interaction throughout the scoring period?

Methods

Research Model

This study aims to examine the changes in the ratings of peer raters over time in evaluating the presentation skills of 7th-grade students. For this purpose, as one of the quantitative research approaches, the descriptive research model was used (Şata, 2020).

The Study Group

The present study was conducted in the first semester of the 2021-2022 education year in Turkiye. The study group of the research consisted of 7th-grade students (N=28) studying in a private school in Ankara, Çankaya district. The presentations of the science and technology course prepared by the 7th-grade students in groups of three (i.e., eight groups) were scored by peer raters using rubrics. Since each presentation was made to a group of three people, the total number of peer raters was 25, and 8 groups carried out the scoring.

Instruments

In peer assessment of presentations, students use an analytical rubric developed by the researchers (see Appendix 1). While developing the rubric, researchers determined the criteria for assessing students' presentation skills by reviewing the relevant literature. The scale was developed as an analytical rubric, preferable to the holistic one since they provide more detailed feedback on student performance. One of the most important advantages of analytical rubrics is that they are better than holistic rubrics in providing both intra-rater and inter-rater reliability in the assessment process (Andrade, 2005), which is the main reason that led us to adopt them in the present study. The response categories were initially developed as five but were subsequently reduced to four. Respectively, content, coherence, use of material, communication, and use of time. To provide evidence for the reliability and validity of the measurements obtained from the analytical rubric, content validity rates and Kendall Tau coefficient were calculated through expert opinions. In line with the opinions of field experts (6 people) and assessment and evaluation experts (3 people), the rating of the rubric was also reduced from five to four in the final version. To measure the reliability, Kendall tau was calculated to measure the relation between the scores of the two randomly selected groups. It was determined that the rubric had an acceptable reliability score ($r = .652$; $p < .001$).

Data Collection

Since the present research attempts to examine rater change over time, eight groups made presentations, two groups each day, in four days. In the evaluation process of the presentations, the analytical rubric was used, and the criteria of the measurement tool were introduced to the students one by one before they made their presentations to ensure the consistency of the measurements. The scores of the peer raters at the end of each presentation were collected. The next group made their presentation after having a short break. Students followed the same procedure on each of the four days.

Data Analysis

Statistical analysis of the research was carried out using the multi-facet Rasch model, which is a member of the Rasch model family. Many-Facets Rasch Analysis (Many-Facets Rasch Model) is a useful measurement model since it considers all sources of variability that affect individuals' performance or skill levels (Baird et al., 2013; Kim et al., 2012; Linacre, 2017). In addition, it provides the opportunity to examine the interaction among the variability sources (Kassim & Noor, 2007). The simultaneous analysis of both two- and multi-category measurements increases the applicability of the multi-facet Rasch model. In the current study, the Many-Facets Rasch model was used since it aimed to examine both the main effects and the common interactions among sources of variability in the peer-assessment process. In this study, standardized differences (Signed Area Index, SAI) obtained from multi-facet Rasch analysis and interaction terms were used to examine the change of peer raters over time.

Measurements at different times were estimated as separate models to explore standardized differences. Estimated logit values for each model were divided by standard errors, and standardized values were obtained. In this study, since student presentations were different in the measurements made at different times, the mean score of the estimations of the raters was modeled to be zero to eliminate the influence of this difference. As such, the relative change in the strictness or leniency behaviors of the raters could be examined in the scoring that takes place at different times. For this reason, student presentations (groups) were handled as non-center in the research. For the standardized differences, the first measurement baseline time was taken, the measurements at other times were compared with the baseline time, and score deviations (SAI) were calculated. The SAI value has been standardized by the formula given below.

$$Z_{SAI_{Difference}} = \frac{M_c - M_b}{\sqrt{SE^2_{M_c} + SE^2_{M_b}}} \qquad (1)$$

Here, Mc corresponds to rater strictness or leniency compared to baseline time, while Mb corresponds to rater strictness and leniency at baseline time. The two values in the denominator represent the squares of the standard errors of the rater severity or leniency values at two different times. If the $Z_{SAI_{Difference}}$ value is calculated from two different times out of the range of ±1.96 values, it indicates a statistically significant difference (Raju, 1990). When the direction of the facets is positive, positive $Z_{SAI_{Difference}}$ values are interpreted as the rater becomes lenient, whereas negative $Z_{SAI_{Difference}}$ values are interpreted as showing severity over time (Börkan, 2017).

Another index used to examine the change of peer raters' score over time is the term interaction (Wolfe et al., 2007). Time is added to the model as a dummy variable for the interaction index, and the interactions between the rater and the time variable are examined (Börkan, 2017).

Since the model data fit needs to be achieved in order to produce consistent and unbiased estimations in the multi-facet Rasch model, standardized residuals were examined. 1% of the standardized residual values were in the range of ±3, and 5% were in the range of ±2, indicating a sufficient model-data fit (Linacre, 2017). In the current study, 3 (0.03%) of the standardized residual values were found to be in the range of ±3, and 37 (3.70%) of them were found to be in the range of ±2 (total number of observations 25 raters x 8 groups x 4 criteria = 800). These results indicated that the model-data fit was acceptable, and the estimations were unbiased and consistent.

Results

Within the scope of the present study, the change in the scores of the peer raters over time was examined first. As the first day is determined as a baseline, the variation of the other days from the first day was

examined. The logit values and standard differences obtained from the four measurements are given in Table 1.

Table 1

The Change of Peer Raters' Ratings Over Time

	Day 1		**Day 2**		**Day 3**		**Day 4**		**Z SAIdifference**		
Rater	**Logit**	**SH**	**Logit**	**SH**	**Logit**	**SH**	**Logit**	**SH**	**2-1**	**3-1**	**4-1**
1	-1.10	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.53	1.38	1.08
2	-0.68	0.65	-0.59	0.70	-0.26	0.68	0.77	0.72	0.09	0.45	1.49
3	-0.25	0.65	-1.06	0.67	-0.26	0.68	-0.50	0.61	-0.87	-0.01	-0.28
4	0.17	0.65	-1.06	0.67	0.20	0.68	-0.50	0.61	-1.32	0.03	-0.75
5	1.57	0.74	1.59	0.77	0.67	0.68	0.77	0.72	0.02	-0.90	-0.77
6	1.57	0.74	0.46	0.74	1.15	0.70	0.77	0.72	-1.06	-0.41	-0.77
7	1.06	0.69	-0.08	0.73	0.67	0.68	-0.12	0.63	-1.13	-0.40	-1.26
8	1.57	0.74	-0.08	0.73	1.15	0.70	0.77	0.72	-1.59	-0.41	-0.77
9	1.57	0.74	-0.08	0.73	-2.60	0.68	-2.29	0.61	-1.59	**-4.15**	**-4.02**
10	-0.25	0.65	-0.59	0.70	-0.26	0.68	0.29	0.66	-0.36	-0.01	0.58
11	-1.10	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.53	1.38	1.08
12	-1.52	0.65	-0.59	0.70	-0.26	0.68	0.29	0.66	0.97	1.34	1.95
13	-0.68	0.65	-0.59	0.70	0.20	0.68	-0.12	0.63	0.09	0.94	0.62
14	0.17	0.65	-1.06	0.67	-0.73	0.69	2.24	1.09	-1.32	-0.95	1.63
15	-1.52	0.65	-0.59	0.70	-0.26	0.68	1.36	0.83	0.97	1.34	**2.73**
16	-1.52	0.65	-0.59	0.70	0.20	0.68	2.24	1.09	0.97	1.83	**2.96**
17	-0.68	0.65	3.15	1.09	-1.67	0.69	-0.87	0.60	**3.02**	-1.04	-0.21
18	-0.25	0.65	-0.59	0.70	1.15	0.70	-1.22	0.60	-0.36	1.47	-1.10
19	-1.52	0.65	-0.59	0.70	-0.26	0.68	-0.87	0.60	0.97	1.34	0.73
20	-0.25	0.65	1.01	0.75	-1.20	0.69	-0.87	0.60	1.27	-1.00	-0.70
21	1.57	0.74	0.46	0.74	1.15	0.70	-1.93	0.60	-1.06	-0.41	**-3.67**
22	1.57	0.74	1.59	0.77	0.67	0.68	0.29	0.66	0.02	-0.90	-1.29
23	0.17	0.65	-1.06	0.67	0.67	0.68	0.77	0.72	-1.32	0.53	0.62
24	-0.25	0.65	0.46	0.74	-0.26	0.68	-0.87	0.60	0.72	-0.01	-0.70
25	0.61	0.66	1.59	0.77	-0.26	0.68	-0.12	0.63	0.97	-0.92	-0.80
Ort.	0.00		0.00		0.00		0.00		-0.03	0.02	-0.07
SD	1.11		1.06		0.90		1.10		1.14	1.28	1.67

Those with thick font sizes represent raters who performed rater drift statistically.

When Table 1 was examined, it was seen that the scores of 25 peer raters from day 1 to day 2 decreased by -0.03 points on average. On the third day, it was seen that the scores increased by 0.02 on average and decreased by 0.07 on the last day. It was stated that for a significant group-level rater severity or leniency, $Z_{SAI_{Difference}}$ should be 0.50 and above between two-time measures (Swaminathan & Rogers, 1990). When the average point of the other three-time measurements compared to the baseline time was examined, it was found to be close to zero, and all times had almost the same level of severity or leniency. Although there was no significant rater drift at the group level, it was revealed that some raters at the individual level tended to be more severe or lenient over time. For example, on day 2, rater 17 displayed a more lenient behavior than on the first day. As a result, it was presented that a few raters at the individual level made more severe or lenient ratings over time, but there was no significant rater drift at the group level.

In addition to using standard differences in determining rater drift, rater drift was examined with the interaction term by examining common effects. Accordingly, within the scope of the present study, the time variable was included in the model, the rater x time interactions were examined, and the findings are given in Table 2.

Table 2

Rater X Time Interactions

	Day 1		Day 2		Day 3		Day 4		$I_{difference}$		
Rater	**Bias**	**SH**	**Bias**	**SH**	**Bias**	**SH**	**Bias**	**SH**	**2-1**	**3-1**	**4-1**
1	-0.72	0.59	-0.09	0.60	0.37	0.61	0.48	0.63	0.75	1.28	1.39
2	-0.55	0.60	-0.28	0.60	-0.19	0.60	1.17	0.71	0.32	0.42	1.85
3	0.10	0.61	-0.36	0.60	0.09	0.60	0.18	0.61	-0.54	-0.01	0.09
4	0.28	0.62	-0.55	0.60	0.28	0.61	0.00	0.61	-0.96	0.00	-0.32
5	0.35	0.72	0.23	0.71	-0.56	0.63	0.04	0.71	-0.12	-0.95	-0.31
6	0.46	0.72	-0.54	0.63	-0.03	0.66	0.16	0.71	-1.05	-0.50	-0.30
7	0.52	0.67	-0.39	0.61	0.08	0.63	-0.19	0.63	-1.00	-0.48	-0.77
8	0.57	0.72	-0.81	0.61	0.08	0.66	0.27	0.71	-1.46	-0.50	-0.30
9	2.12	0.72	0.73	0.61	-1.32	0.59	-1.25	0.59	-1.47	**-3.70**	**-3.62**
10	-0.18	0.61	-0.28	0.60	-0.19	0.60	0.70	0.66	-0.12	-0.01	0.98
11	-0.72	0.59	-0.09	0.60	0.37	0.61	0.48	0.63	0.75	1.28	1.39
12	-0.98	0.59	0.00	0.60	0.09	0.60	0.98	0.66	1.16	1.27	2.21
13	-0.45	0.60	-0.19	0.60	0.28	0.61	0.38	0.63	0.31	0.85	0.95
14	0.00	0.62	-0.83	0.60	-0.74	0.60	2.40	1.07	-0.96	-0.86	1.94
15	-1.16	0.59	-0.19	0.60	-0.09	0.60	1.84	0.81	1.15	1.27	**2.99**
16	-1.35	0.59	-0.37	0.60	0.09	0.61	2.50	1.07	1.16	1.70	**3.15**
17	-0.45	0.60	2.89	1.07	-1.16	0.59	-0.37	0.60	**2.72**	-0.84	0.09
18	-0.09	0.61	-0.19	0.60	1.08	0.66	-0.74	0.60	-0.12	1.30	-0.76
19	-0.70	0.59	0.27	0.60	0.36	0.60	0.08	0.60	1.15	1.26	0.93
20	0.00	0.61	1.08	0.66	-0.72	0.59	-0.28	0.60	1.20	-0.85	-0.33
21	1.19	0.72	0.19	0.63	0.70	0.66	-1.82	0.59	-1.05	-0.50	**-3.23**
22	0.46	0.72	0.35	0.71	-0.45	0.63	-0.31	0.66	-0.11	-0.95	-0.79
23	-0.10	0.62	-0.93	0.60	0.28	0.63	0.88	0.71	-0.96	0.43	1.04
24	-0.09	0.61	0.57	0.63	-0.09	0.60	-0.37	0.60	0.75	0.00	-0.33
25	0.10	0.63	0.88	0.71	-0.67	0.60	-0.19	0.63	0.82	-0.89	-0.33
Ort.	-0.06		0.04		-0.08		0.28		0.09	0.00	0.31
SD	0.75		0.79		0.55		0.99		1.06	1.18	1.61

Fixed (all = 0) chi-square: 116.3 d.f.: 100 significance (probability): .130

Variance explained by the interaction (%): 11.24

When Table 2 was examined, it was explored that the rater x time interaction was not statistically significant at the group level ($\chi 2$ (100) = 116.30, p > .05). At the individual level, it was revealed that the scores of some raters at different times showed a drift. With interaction analysis, it was found that the standard differences gave similar results. The fact that the rater drifts, which were statistically significant in standardized differences, were also significant as a result of the interaction analysis indicates that both techniques are powerful in detecting rater drifts. The graph of the t-values for rater time interactions is given in Figure 1.

Figure 1 shows that there are more values outside the ±1.96 range, especially in the fourth time measurement. This finding provides evidence that peer raters' scores may vary over time. There is a smaller change in the scoring times of the 3rd time compared to the other times. The ratings of each rater over time are given in Figure 2.

Figure 1

T-values for Rater x Time Interactions

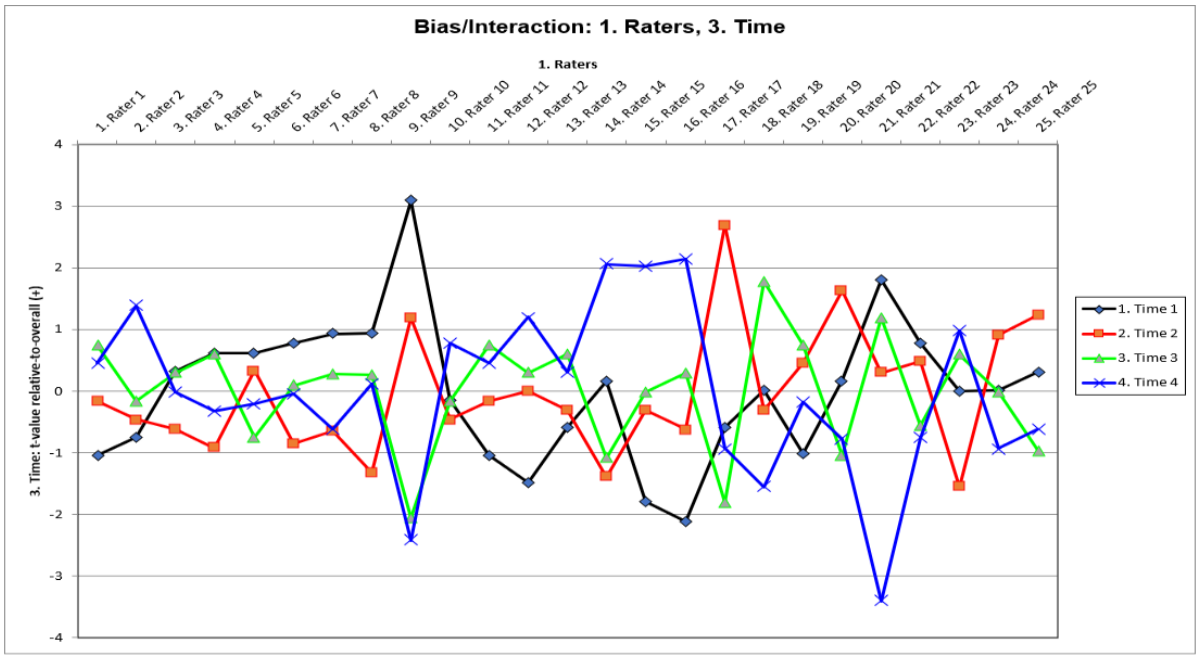
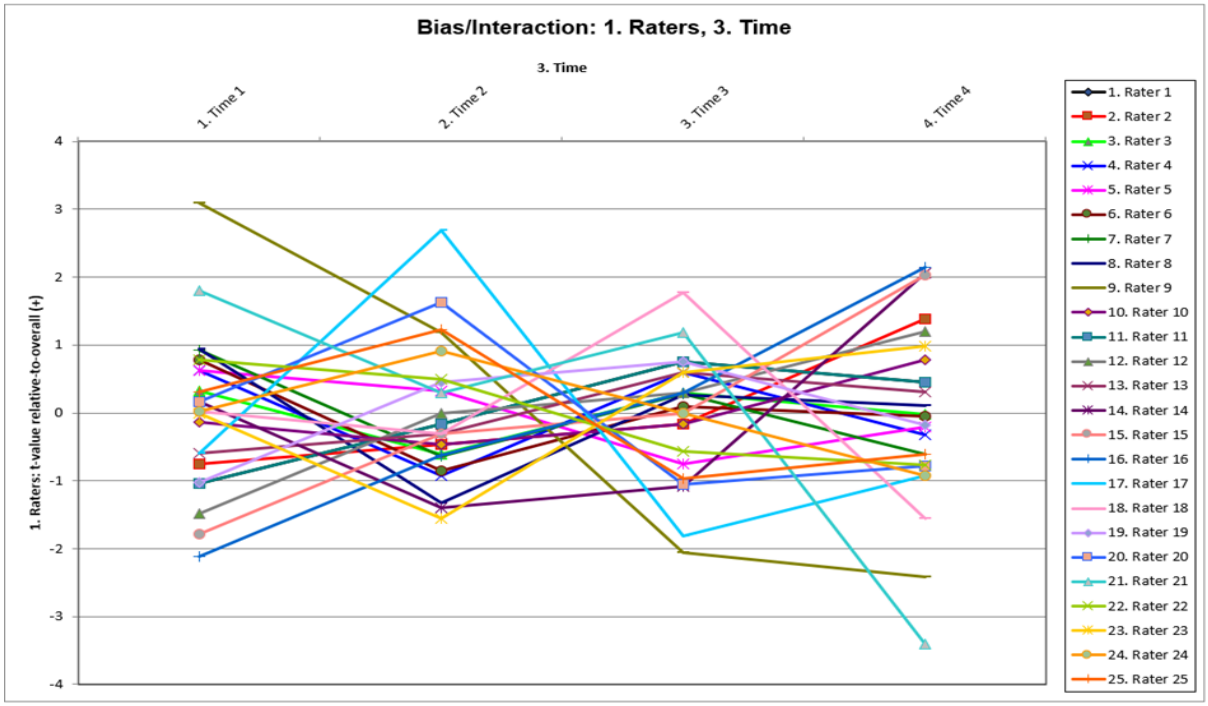


Figure 2

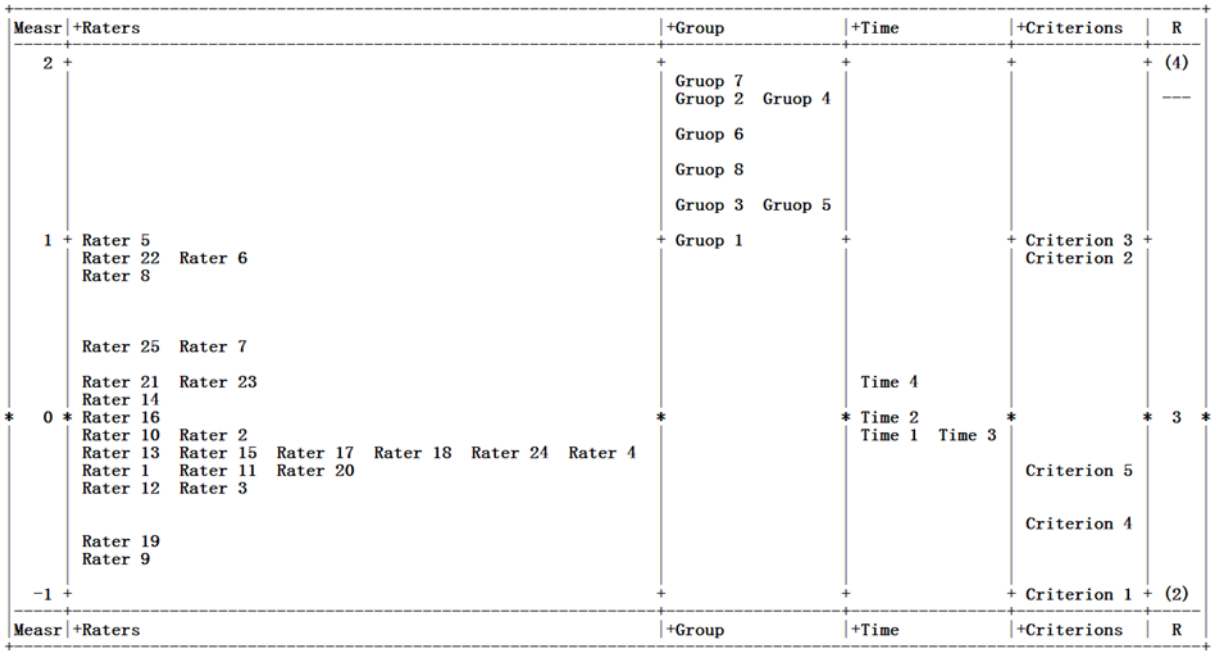
T-values for Each Raters' Ratings Over Time



When Figure 2 was examined, we observed that the t-values for raters 9, 17, and 21 were higher than the critical t-values, as well as rater drift over time. It was determined that the other raters had small rater drifts over time, but there were no statistically significant changes. Finally, the variable map, which allows examining all facets together, is given in Figure 3.

Figure 3

The Variable Map



When Figure 3 is examined, it is seen that the four facets within the scope of the research are placed on a single scale (logit scale). Thus, the four facets can be compared graphically with respect to each other. For example, who is the severity rater or who is the most successful group can be seen simultaneously.

Discussion and Conclusion

It is stated that peer assessment, which has increased in importance with alternative approaches, can enhance students' motivation in the learning environment (Topping, 2009) and contribute to their development of in-depth thinking and problem-solving skills (Patri, 2002). However, there is limited evidence for the validity of peer-reviewed scores. Therefore, it is considered necessary to carry out studies on the validity of the scores obtained from peer assessment, which is gaining importance so far.

In this study, rater severity drift, one of the rater effects, was examined in evaluating students' presentations that were spread through time by their peers. In the analyses made in peer assessment applications, it was tried to observe whether there was peer rater severity drift in the following three scoring days, compared with the first day, which was grounded based on the first day of scoring. In the study conducted with middle school 7th-grade students, peer assessment lasted four days, and peer raters used rubrics to evaluate the presentations. Accordingly, whether the peer raters drifted their peer severity at the individual or group level during the presentation was analyzed with the help of MFRM analysis. Standardized differences and interaction term (rater x time) approaches were used to determine rater drift at the group level. The results obtained from both approaches showed similar characteristics. According to the results, it was observed that the scores of the peer raters at the group level became severe from “Day 1” to “Day 2”, and their scores were more lenient on “Day 3” and became severe again on “Day 4”, yet these differences were not significant at the group level. This finding indicates that the students did not make a biased scoring while evaluating the group performance according to the groups. Although there was no significant rater drift at the group level, it was observed that some students at the individual level showed rater drifts over time. In determining rater drift at the individual level, the raters were compared among themselves using the standard differences and common effects (rater x time) approaches. The results obtained from both approaches showed similar characteristics. Accordingly, three of the students participating in the study demonstrated a rater drift over time. Two

students (numbered 9 and 21) scored more severely over time, while one student (numbered 17) scored more leniently. Finally, when the variability map enabled us to examine all facets together, it was found that the most lenient rater was 5 and the most severe rater was 9. In the presentations, it was revealed that the most successful group was the 7th and the most unsuccessful group was the 1st group, that the presentations were consistently ranked by the raters according to their qualities, and it was determined that the drifts on other days were close to each other, except for the 4th day, in time measurements. According to the change map, it was also found out that the criterion that the groups had the least difficulty with was the 3rd criterion, "Material Use", and the criterion they had the most difficulty with was the 1st criterion, "Content".

The study group was an experienced one in the use of peer assessment approach and rubrics in the educational environment. The reason why the students did not make rater drift on a group basis may be that they included peer assessment practices in the learning process and therefore, they were experienced in this regard. It is stated in the related literature that rater education reduces rater effect (Hauenstein & McCusker, 2017). Another reason why students did not have rater drift at the group level may be that the presentation scores were spread over a short period of time (4 days). Harik et al. (2009) have stated that in studies whose scoring was done within days or weeks, rater drift could be at a minimum level when compared to studies whose scoring was spread over months or years.

Although there was no rater drift at the group level, it was observed that a small portion of the students (3 students) made more lenient ratings over time. Previous literature has different findings on this issue. While one study has shown that leniency increased over time (Lunz & Stahl, 1990), four studies have shown increasing severity (Congdon & McQueen, 2000; McLaughlin et al., 2009; Myford, 1991; Pinot de Moira et al., 2002), and two other studies have demonstrated positive and negative drift for a small proportion of their raters (Börkan, 2017; Myford & Wolfe, 2009).

Analyzing rater drift at the individual level, we found that two students displayed severity drift and three students displayed lenient drift. In the literature, it is seen that the leniency of raters, which differs in the peer assessment process, is quite common (Erman-Aslanoğlu et al., 2020; Farrokhi et al., 2012). Considering both standard differences and interaction/bias analysis, we can conclude that both methods can be used to detect the rater drift of the same individuals separately. Therefore, it will suffice to choose one method for future research. Moreover, considering that interaction analyses are systemic errors, they have a negative impact on the validity of measurements obtained from these individuals (Messick, 1996). As a result, evidence was provided for the reliability and validity of the measurements at both the group and individual levels in the peer assessment process, and it was determined that some peers had an effect on the validity of the measurements at the individual level. However, there was no statistically significant effect on the validity and reliability of the measurements at the group level. The exclusion of students with rater drift from the scoring will, therefore, contribute to the reliability and validity of the measurements if the evaluation of the students is crucial.

This research is limited to 7th-grade level and oral presentation skills. Researchers can investigate the effect of rater drift on peer ratings over time by conducting similar studies at different grades and with different skills. Researchers can also examine the effect of peer drift in terms of students who actively use the peer assessment approach in the teaching environment and who do not use this assessment approach. To reduce rater drifts that occur over time in peer assessment, rater training can be designed. Thus, it is expected that the measurements obtained in the performance assessment will contribute to the reliability and validity.

Declarations

Author Contribution: Aslıhan Erman Aslanoğlu: Theoretical framework, literature review, methodology, data collection, data analysis, discussion, writing the original draft and review & editing. Mehmet Şata: Theoretical framework, methodology, data analysis, discussion and review & editing.

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

Ethical Approval: The study was approved by the Ufuk University Ethics Committee (Research code: 2021-49, dated 09.06.2021)

References

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching*, *53*(1), 27-31. https://doi.org/10.3200/CTCH.53.1.27-31

Azarnoosh, M. (2013). Peer assessment in an EFL context: attitudes and friendship bias. *Language Testing in Asia*, *3*(1), 1-10. https://doi.org/10.1186/2229-0443-3-11

Baird, J. A., Hayes, M., Johnson, R., Johnson, S., & Lamprianou, I. (2013). *Marker effects and examination reliability. A Comparative exploration from the perspectives of generalisability theory, Rash model and multilevel modelling*. Oxford University Centre for Educational Assessment. https://core.ac.uk/download/pdf/15171449.pdf

Batmaz, H., Türk, N., Kaya, A., & Yıldırım, M. (2022). Cyberbullying and cyber victimization: examining mediating roles of empathy and resilience. *Current Psychology*, 1-11. https://doi.org/10.1007/s12144-022-04134-3

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. McGraw-Hill. http://www.mcgraw-hill.co.uk/html/0335212972.html

Börkan, B. (2017). Rater severity drift in peer assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(4), 469-489. https://doi.org/10.21031/epod.328119

Bushell, G. (2006). Moderation of peer assessment in group projects. *Assessment and Evaluation in Higher Education*, *31*(1), 91–108. https://doi.org/10.1080/02602930500262395

Congdon, P. J., & McQueen, J. (2000). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement, 37*(2), 163-178. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2000.tb01081.x

Dochy, F. (2001). A new assessment era: different needs, new challenges. *Learning and Instruction*, *10*(1), 11-20. https://doi.org/10.1016/S0959-4752(00)00022-0

Donnon, T., McIlwrick, J., & Woloschuk, W. (2013). Investigating the reliability and validity of self and peer assessment to measure medical students' professional competencies. *Creative Education*, *4*(6A), 23-28. http://dx.doi.org/10.4236/ce.2013.46A005

Erman Aslanoğlu, A., Karakaya, İ., & Şata, M. (2020). Evaluation of university students' rating behaviors in self and peer rating process via many facet rasch model. *Eurasian Journal of Educational Research*, 20(89), 25-46. https://dergipark.org.tr/en/pub/ejer/issue/57497/815802

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. https://doi.org/10.2307/1170785

Farrokhi, F., Esfandiari, R. ve Dalili, M.V. (2011). Applying the Many-Facet Rasch Model to detect centrality in self-Assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal 15 (Innovation and Pedagogy for Lifelong Learning)*, 70-77. http://www.idosi.org/wasj/wasj15(IPLL)11/12.pdf

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet Rasch measurement of differential rater severity/leniency in three types of assessment. *JALT Journal*, 34(1), 79-101.

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International Journal of Science Education, 25*(12), 1509–1528. https://doi.org/10.1080/0950069022000038268

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. 2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement, 46*(1), 43-58. https://doi.org/10.1111/j.1745-3984.2009.01068.x

Hauenstein, N. M. A. & McCusker, M. E**.** (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment, 25*, 253-266. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2009.01068.x

Humphris GM, & Kaney S. (2001). Examiner fatigue in communication skills objective structured clinical examinations. *Medical Education*, *35*(5), 444-449. https://doi.org/10.1046/j.1365-2923.2001.00893.x

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kaya, A., Türk, N., Batmaz, H., & Griffiths, M. D. (2023). Online gaming addiction and basic psychological needs among adolescents: the mediating roles of meaning in life and responsibility. *International Journal of Mental Health and Addiction*, 1-25. https://doi.org/10.1007/s11469-022-00994-9

Kassim, A.N.L. (2007, June 14-16). *Exploring rater judging behaviour using the many-facet Rasch model* [Conference Session]. The Second Biennial International Conference on Teaching and Learning of English

in Asia: Exploring New Frontiers (TELiA2), Universiti Utara, Malaysia. http://repo.uum.edu.my/id/eprint/3212/

Kim, Y., Park, I., & Kang, M. (2012). Examining rater effects of the TGMD-2 on children with intellectual disability. *Adapted Physical Activity Quarterly, 29*(4), 346-365. https://doi.org/10.1123/apaq.29.4.346

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training?. *Assessing Writing, 12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Kubiszyn, T., & Borich, G. (2013). *Educational testing and measurement: Classroom application and practice* (10th ed.). John Wiley & Sons. https://l24.im/jV6yYCJ

Kutlu, Ö., Doğan, C. D., & Karaya, İ. (2014). *Öğrenci başarısının belirlenmesi: Performansa ve portfolyoya dayalı durum belirleme* [Determining student success: Assessment based on performance and portfolio]. Pegem. https://l24.im/k5cn

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399-418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

Linacre, J.M. (2017). *A user's guide to FACETS: Rasch-model computer programs.* MESA Press.

Liu, H., & Brantmeier, C. (2019). "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System, 80*, 60-72. https://doi.org/10.1016/j.system.2018.10.013

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*(4), 425 444. https://psycnet.apa.org/doi/10.1177/016327879001300405

McLaughlin K, Ainslie M, Coderre S, Wright B & Violato C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Medical Education, 43*(10), 989–992. https://doi.org/10.1111/j.1365-2923.2009.03438.x

Messick, S. (1994). Alternative modes of assessment, uniform standards of validity. *ETS Research Report Series, 2*, 1-22. https://doi.org/10.1002/j.2333-8504.1994.tb01634.x

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). National Center for Education Statistics.

Moore, B. B. (2009). *A consideration of rater effects and rater design via signal detection theory* (Publication No. 3373803). [Doctoral dissertation, Columbia University]. ProQuest Dissertations & Theses Global.

Mulqueen C., Baker D. & Dismukes R.K. (2000, April). *Using multifacet rasch analysis to examine the effectiveness of rater training* [Conference Session]. 15th Annual Meeting of the Society for Industrial/Organizational Psychology, New Orleans, LA. https://www.air.org/sites/default/files/2021-06/multifacet_rasch_0.pdf

Myford, C. M. (1991, April 3-7). *Judging acting ability: The transition from notice to expert* [Conference Session]. Annual Meetin of the American Educational Research Association, Chicago IL. https://files.eric.ed.gov/fulltext/ED333032.pdf

Myford, C. M., & Wolfe, E. M. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Aplied Measurement, 4*(4), 386-422. https://psycnet.apa.org/record/2003-09517-007

Myford, C. M., & Wolfe, E. M. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale use. *Journal of Educational Measurement, 46*(4), 371-389. https://psycnet.apa.org/doi/10.1111/j.1745-3984.2009.00088.x

Oosterhof, A. (2003). *Developing and using classroom assessment* (3th ed.). Merrill/Prentice Hall. https://l24.im/OCKvkg2

Orsmond P, Merry S, Reiling K (2000) The use of student-derived marking criteria in peer and self-assessment. *Assessment &Evaluation Higher Education, 25*(1), 21–38. https://doi.org/10.1080/02602930050025006

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing, 19*(2), 109–131. https://doi.org/10.1191/0265532202lt224oa

Pinot de Moira, A., Massey, C., Baird, J., & Morrissey, M. (2002). Marking consistency over time. *Research in Education, 67*(1), 79–87. https://doi.org/10.7227/RIE.67.8

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*(2), 197-207. https://psycnet.apa.org/doi/10.1177/014662169001400208

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1990.tb00754.x

Şata. M. (2020a). Quantitative research approaches [Nicel araştırma yaklaşımları]. In Oğuz. E. (Ed.). *Research methods in education [Eğitimde araştırma yöntemleri*] (p. 77-98). Eğiten Kitap publishing.

Topping, K. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. S. Segers, Dochy,R., & E. C. Cascallar (Eds.), *In optimising new modes of assessment: In search of qualities and standards* (pp. 55-87). Springer Dordrecht. https://doi.org/10.1007/0-306-48125-1

Topping, K. (2009). Peer assessment. *Theory Into Practice*, *48*(1), 20-27. https://doi.org/10.1080/00405840802577569

Topping, K. (2017). Peer assessment: learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology, 1(*1), 1-17. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.007

Wolfe, E. W., Myford, C. M., Engelhard Jr. G., & Manalo, J. R. (2007). *Monitoring reader performance and DRIFT in the AP® English literature and composition examination using benchmark essays.* College Board. https://files.eric.ed.gov/fulltext/ED561038.pdf

Yurdabakan, İ., & Cihanoğlu, M. O. (2009). The effects of cooperative reading and composition technique with the applications of self and peer assessment on the levels of achivement, attitude and strategy use. *Dokuz Eylul University The Journal of Graduate School of Social Sciences, 11*(4), 105-123. https://l24.im/VbHg

Appendix

Data Collection Tool

Criteria	1	2	3	4
Content	Topic is irrelevant and focused; presentation contains multiple fact errors.	Topic should be more focused and relevant; presentation contains some fact errors or omissions.	Topic is adequately focused and relevant; major facts are accurate and generally complete.	Topic is tightly focused and relevant; presentation contains accurate information with no fact errors.
Coherence	Ideas are not presented in proper order; transition is lacking between major ideas; several parts of the presentation are wordy or unclear.	Some ideas are not presented in proper order; transition markers are needed between some ideas; some parts of the presentation are wordy or unclear.	Most ideas are in logical order with adequate transitions between most major ideas; presentation is generally clear and understandable.	Ideas are presented in logical order with effective transitions between major ideas; presentation is clear and concise.
Use of Material	No material is used in the presentation.	Presentation is supported with a relevant material.	Presentation is supported with 2 different relevant materials.	Presentation is supported with 3 different relevant materials.
Communication	Inadequate voicing or energy, too slow or too fast pacing, poor pronunciation, distracting gestures or posture, unprofessional appearance, and visual aids poorly are used.	Neither adequate nor inadequate voicing and energy; slow or fast pacing; some distracting gestures or posture; adequate appearance; few visual aids are used.	Adequate voicing and energy; generally good pacing and intonation; few or no distracting gestures; professional appearance; adequate visual aids are used.	Proper voicing and energy; good pacing and intonation; no distracting gestures; professional appearance; effective and adequate visual aids are used.
Use of Time	The presentation exceeds or lags behind the time limit.	The presentation does not comply with the time limit (+/- 3).	The presentation does not comply with the time limit (+/- 2).	The presentation is completed on time.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 76-94

Comparison of Methods Used in Detection of DIF in Cognitive Diagnostic Models with Traditional Methods: Applications in TIMSS 2011*

Büşra EREN** Tuba GÜNDÜZ*** Şeref TAN****

Abstract

This study aims to compare the Wald test and likelihood ratio test (LRT) approaches with Classical Test Theory (CTT) and Item Response Theory (IRT) based differential item functioning (DIF) detection methods in the context of cognitive diagnostic models (CDMs), using the TIMSS 2011 dataset as a retrofitting study. CDMs, which have a significant potential when determining the DIF and their contribution to validity, can give confidence under the strong methodological background condition is met. Therefore, it is hoped that this study will contribute to the literature to ensure the correct usage of CDMs and evaluate the compatibility of these new approaches with traditional methods. According to the analysis results, thirty-one items showed differences between the cognitive diagnosis assessments and the traditional methods. The item with the largest DIF was found in the Raju Unsigned Area Measures technique in IRT, whereas the item with the lowest DIF was found in the Wald test technique developed for CDMs. In general, the analyses show that methods not based on CDMs detect more items with DIF, but the Wald test and LRT methods based on CDMs detect fewer items with DIF. This study conducted DIF analyses to determine the test's psychometric properties within the framework of CDMs rather than the source of the bias. Researchers can take the study one step further and make more specific assessments about the items' bias regarding the test structure, test scope, and subgroups. In addition, DIF analyses in this study were carried out using only the gender variable, and researchers can use different variables to conduct studies specific to their purpose.

Keywords: Cognitive diagnosis models, large scale assessment, differential item functioning

Introduction

Cognitive Diagnostic Models (CDMs) are psychometric models that provide detailed information about examinees' mastery of interrelated but separable attributes (Hou et al., 2014). Rather than dealing with students' positions on a continuous latent variable as Item Response Theory (IRT) does, CDMs predict a profile of categorical latent attributes. The term "attribute" is used to refer to latent variables in the study because latent variables assume that the items in the measurement tool may be related to one or more latent variables, which are referred to by various names in the literature, such as ability, and skill (Paulsen et al., 2020; Ravand & Baghaei, 2019).

CDMs, which provide examinees with finer-grained diagnostic information, enable them to be classified according to their mastery profiles (DiBello & Stout, 2007). In this classification, the correct response to the item indicates that the student has the necessary attributes, represented by "1" in the Q-matrix entries. Otherwise, these entries are "0" (Rupp et al., 2010). This matrix, which is essential in

*The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Şeref TAN and prepared by Büşra EREN.

** Instructor, National Defence University, Department of Educational Sciences, Balikesir-Türkiye, busra_karaduman@yahoo.com, ORCID ID: 0000-0001-7565-1025
*** Res. Assist. PhD., Mugla Sitki Kocman University, Faculty of Education, Mugla-Türkiye, tuba.karacan@yahoo.com, ORCID ID: 0000-0002-0921-9290
**** Prof. Dr., Ankara-Türkiye, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 76-94. https://doi.org/10.21031/epod.1218144

Received: 12.12.2022
Accepted: 6.03.2023

determining the profiles of students regarding the attributes that do plan to measure with the test and which is confirmatory, is a common point of CDMs. It maps the attributes required by the items in a multidimensional way by placing them in rows and the attributes in columns, using a simple or complex load structure (de la Torre & Minchen, 2014; Rupp et al., 2010). The validity of the findings gathered from the students' responses increases when the items and attributes in the matrix correctly match within the framework of the relevant structure (Ravand & Baghaei, 2019). Therefore, identifying the Q-matrix used in CDMs becomes essential in testing development when considering its accuracy and design (Kang et al., 2018). When this step omits, the studies' findings indicate biases in item parameters and problems in student classification (de la Torre, 2008; de la Torre & Chiu, 2016). Bias in the parameters is among the factors affecting the validity. It may occur when the scores of students in different subgroups contain systematic errors (Camilli & Shepard, 1994).

It is stated in the literature that many important statistical routines are needed to ensure appropriate uses and interpretations and to unlock the potentials of CDMs, such as the procedure for detecting differential item functioning (DIF), which can be used to determine item parameter bias (Ma et al., 2021; Paulsen et al., 2020). DIF has been described traditionally as "the probability of students with the same total test score or ability level but in different groups to correctly respond to an item when the variable is unrelated to the construct of interest" (Hou et al., 2014). For example, suppose an item has a systematic advantage favouring the female group. The item might be biased since the item response function differs between the female and male groups. A problem will arise regarding the validity of scores obtained from the test since different properties are mixed with the property to be measured. Researchers should identify and examine biased items to eliminate the problem and perform proper measurement procedures (Lee et al., 2021).

DIF is as essential in CDMs as it is in traditional approaches. "Traditional approaches make rankings at the latent ability level, while CDMs focus on the change in correct response probability regarding the responses given to an item by students in various groups but with similar attribute mastery profiles" according to the difference between the two types. In other words, DIF is defined according to CDMs as "an effect in which the probability of answering an item is different correctly for students with the same attribute mastery profile but from different observed groups" (Hou et al., 2014).

CDMs, a multidimensional model that has been increasingly popular in recent years, are used to obtain diagnostic information about students' strengths and weaknesses. This information, along with feedback opportunities for teachers and programs, provides students with opportunities for individualized learning support that compensates for learning deficiencies. In addition to this contribution, DIF detection, one of the most important statistical routines for ensuring proper usage and interpretation, appears to be a helpful method, mainly when dealing with the issue of validity, which is a problem with traditional methods (Akbay, 2021). Therefore, detecting DIF has become a standard procedure in psychometric analyses. The presence of items with DIF can worsen the predictions of attributes (Paulsen et al., 2020) and disturb the attribute profiles, causing problems in comparing latent classes between groups (Hou et al., 2014). DIF analysis is also necessary to examine parameter or configural invariance (Zumbo, 2007). According to attribute profiles, item responses that must independently condition are considered invariant. As a result, DIF analysis is critical for determining whether attribute-item interactions between groups are invariant (Hou et al., 2014).

There is little research on determining DIF in CDMs in the literature. Milewski and Baron (2002) applied DIF to individual skill performance and compared the results of four DIF methods without considering item biases. Zhang (2006) compared traditional methods limited to uniform DIF (MH and SIBTEST) at the level of attribute and item in determining DIF and took into account different simulation conditions using deterministic noisy “and” gate (DINA). When the findings were examined, when the conditions related to the test scores and the attribute profiles were taken into account, it was seen that the matching in the attribute profiles resulted in lower Type I error and higher power rates compared to the test scores, but both methods showed poor performance. Li (2008) extended High-Order DINA (HO-DINA), which was developed by de la Torre and Douglas (2004), to examine DIF and differential attribute functioning (DAF) simultaneously. The new approach used the MCMC algorithm, including Gibbs sampling, to

estimate and compare the model with the traditional MH technique regarding Type I error and power rates under different conditions. In addition, the presented new approach was examined in real-life conditions using a mathematics test. In their simulation study, Hou et al. (2014) developed a new technique (Wald test) to analyze uniform and non-uniform DIF in CDMs. With a simulation study, Liu et al. (2019) investigated the performance of the Wald test in determining DIF using various covariance matrices. To determine DIF in CDM, Hou et al. (2020) utilized the Wald test formulations. The performances of the items in the real dataset were investigated under various simulations, and the compatibility of the attributes' classifications was evaluated when saturated and reduced models were used. In the CTT, IRT, and CDM framework, Akbay (2021) investigated the test's psychometric properties using DIF determination methods (i.e., MH, Raju area measures, and Wald test for DIF). DIF flagging patterns of three different DIF detection methods were observed when real data from a large-scale assessment (TEOG) were retrofitted. The data was collected using Booklets A and B, and DIF analyses were conducted in subgroups based on gender and booklet-type variables. Finally, the studies of Ma et al. (2021) changed the assumptions of the multi-group G-DINA model (MG G-DINA). They developed the MG-G-DINA model for DIF detection to reveal that students in different groups could use the same or different attributes in various ways and compared the performance of this model with the likelihood ratio test (LRT) and Wald test.

Even though there are methods for determining DIFs, studies in the literature suggest that a more effective approach for estimating DIFs is still worth investigating. Because most CDM research has been constrained to research settings over the last decade, many psychometric questions about DIFs related to these models remain unanswered. These non-diagnostic assessments have been retrofitted into CDMs to give detailed information while being examined with traditional models. These are crucial steps in shifting from single-score reporting to CDMs that provide more thorough feedback. The retrofitting is thought to be useful in determining the DIF in order to provide detailed inferences about the students and to provide appropriate use and interpretation of CDMs in the context of the validity and reliability of the inferences regarding the test scores, given exam investments (Terzi & Sen, 2019). Searching for meaning in an evaluation without making assumptions about validity will not give the promised benefit or have the desired influence on educational policies. Therefore, the importance of performing CDM analyses with large-scale datasets should be emphasized in the literature because the differentiation of the exam language, the differences between cultures, or the differences in demographic variables such as gender cause some changes in students' performance. Due to these changes, it will be important to consider the situations that may affect student performance in examining scores (Asil & Gelbal, 2012, Odabas, 2016).

Considering the contributions mentioned in the literature regarding the determination of DIF and its validity, CDMs, which have significant potential, can give confidence, provided that the methodology is sound. As a result, in this study, the Wald test (Hou et al., 2014; Ma et al., 2021) and LRT (Ma et al., 2021), which are based on the MG G-DINA model used in cognitive diagnostic assessments, and Mantel-Haenszel (MH; Mantel & Haenszel, 1959) and logistic regression (LR; Swaminathan & Rogers, 1990) methods, which are based on CTT, and Lord's $\chi 2$ (Lord, 1980) and Raju's unsigned area measures Raju (1988), which are based on IRT methods were compared by using a large-scale dataset TIMSS (Trends in International Mathematics and Science Study) 2011 to ensure the correct use of CDMs. The approaches' compatibility and DIF's effect on CDM were examined using these comparisons. For this purpose, the existence of many studies showings that items with DIF in the bias analyses performed between gender groups, especially in numerical fields such as mathematics, played an important role in the selection of gender as the DIF variable within the scope of the study.

Since there is no single effective method for detecting DIF, using more than one method in the literature is recommended. For this reason, more than one method was used in the study.

DIF Detection Methods

Along with the traditional DIF detection methods utilized in the study, this section gives a brief explanation of the DIF detection methods employed in CDMs.

Methods based on CTTs

The Mantel-Haenszel (MH) is a non-parametric uniform DIF determination technique, although being an $\chi 2$ technique suitable for items scored as 1- or for correct/incorrect responses. When DIF has established an advantage across the ability distribution in favor of only one group, this is known as uniform DIF (Swaminathan & Rogers, 1990). This technique splits students into focal and reference groups, classifying the observed scores into several categories. The students with the same test scores will also have the same ability level after the comparison of the scores of the individuals in the groups in terms of their probability of answering the items correctly according to these categories.

In the first step, the method calculates the likelihood ratio for ability levels. As stated by Camilli and Shepard (2004), in order to facilitate the interpretation of these values, the standardized ΔMH value is obtained by taking the natural logarithm of the odds ratios obtained by dividing the odds values of the focal and reference groups. When the ΔMH value is compared to determine whether it is positive or negative, a "+" value indicates that the focal group is superior. In contrast, a "-" value indicates that the reference group is superior. Below are the values for the size of the ΔMH effect, according to Dorans and Holland (1993): |ΔMH|<1 No DIF (Level A); 1<|ΔMH|<1.5 moderate DIF (Level B); 1.5>|ΔMH| and large DIF (Level C).

In Logistic Regression (LR), which is one of the methods that can be used for both DIF types, it was stated that the scores of the items were predicted by group membership and total score. (Zumbo, 1999). While the item is the dependent variable in the technique, the independent variables are the reference and focal groups, and the significance of the effect of two different groups on the item scores is examined. To determine the DIF magnitude for this technique, Zumbo and Thomas (1996) proposed an effect size measure (ΔR^2), widely used in the literature. When the values are examined, the acceptable limit values are ΔR^2<0.13 (Negligible DIF level), 0.13<ΔR^2<0.26 (Moderate DIF level), and ΔR^2>0.26 (Large DIF level).

Methods based on IRTs

Lord's $\chi 2$ is a technique used for both types of DIF. In this technique, item parameters (a_i - item slope parameter and b_{ij} - the item threshold parameter) for the reference and focal groups are calculated separately for each group. The differences in the parameters are controlled according to the IRT model, and the response status of the focal and reference groups to the relevant item is taken into account (Camilli & Shepard, 1994). DIF analysis is used to see if these parameters are the same. It may also be used to test a null hypothesis: "There is no difference between the item parameters between the focal and reference groups." The presence of DIF and the size of the existing effect can be examined by looking at the p and $\chi 2$ values obtained (Hasançebi, 2021).

DIF is connected with the existence or absence of the area between the item characteristic curves (ICCs) in several methods in the literature. Lord (1980) stated that DIFs might occur because one of the two groups with the same ability level at all θ levels has a higher chance of answering the item correctly than the other group. When the ICCs of the two groups intersect, he also pointed out that the DIF for the items becomes complicated. Raju (1988), on the other hand, suggested formulas for calculating the area between the estimated ICCs for the focal and reference groups for one-, two-, and three-parameter models (Camilli & Shepard, 1994). One of these formulas, known as Raju's unsigned area measures technique, is frequently used in the literature to determine both uniform and non-uniform DIFs. The presence of the area between the ICCs obtained for the focal and reference groups was linked to DIF in this technique. When there is no specified area between two ICCs, it means that the item does not have a DIF.

The technique of Raju's unsigned area is popular for determining uniform and non-uniform DIFs in the literature. For one, two, and three-parameter models, Raju (1988) provided methods for determining the

area between the item characteristic curves (ICC) generated for the focal and reference groups (Camilli & Shepard, 1994).

Methods based on CDMs

When the literature is examined, it is stated that the Wald test detects DIF in DINA by using multivariate hypothesis tests. The Wald test, when the focal and reference groups are taken into account, is based on an alternative hypothesis that at least one of the item parameters is different between these two groups. This technique estimates the attribute distributions and item parameters for the focal and reference groups with separate calibrations. In the second stage, the null hypothesis regarding the item parameters of the two groups is tested (Hou et al., 2014). Ma et al. (2021) proposed a new multi-group CDM (MG G-DINA), which enables the responses from different groups to be modelled at the same time, to improve the Wald test's performance in detecting DIF by explaining that students in different groups can use the same or different attributes and they compared the Wald test based on this model and the LRT in detecting DIF. More than one group is calibrated simultaneously in the Wald test based on this model.

Likelihood-ratio test (LRT) is another DIF detection technique used in the MG G-DINA model. According to the literature, this approach based on IRT can be applied under MG G-DINA without any substantial changes. Uniform DIF occurs in cognitive diagnostic assessments when an item supports one of the groups in all attribute profiles. Otherwise, it indicates that non-uniform DIF is present. More detailed information on DIF detection methods, such as MG G-DINA and Wald test and LRT may be found in the studies of Ma et al. (2021) and Mehrazmay et al. (2021).

Methods

Data and Participants

The sample of this study comes from the 2011 administration of the International Association for the Evaluation of Educational Achievement's (IEA) Trends in International Mathematics and Science Study (TIMSS). TIMSS is an independent, international cooperative of national educational research institutions and governmental research agencies dedicated to improving education (Mullis et al., 2009). The sample of this study consists of 488 8th-grade students (48.57% female) who participated in TIMSS 2011 from Turkey. Turkish students tested on Booklet 2 were selected for DIF analyses in this study.

Structure of the Q-Matrix

A Q-matrix consisting of thirteen attributes and thirty-one items developed by Sen and Arıcan (2015) was used in the study. In order to determine the qualifications, the researchers examined the "common core government standards (CCSS)" used to improve the quality of mathematics education. The attribute list of four content areas accepted by the CCSS in 2010 was considered. In mathematics education, four doctoral students matched the items with these attributes. At least two doctoral students must agree that the item is related to the attributes in the Q matrix and that thirty-one items are related to thirteen attributes in the Q matrix (Sen & Arıcan, 2015).

Data Analysis

This study compares DIF detection methods based on CDMs with those based on CTT and IRT. For this purpose, within the scope of the study, gender was considered as a variable, and the analyses were carried out over "Reference group (R): Male students" and "Focal group (F): Female students". The assumptions in the study were examined before proceeding with DIF analyses based on IRT. The two-parameter logistic model (2PLM), which had a considerably better fit, was used for IRT-based DIF analyses. Before proceeding to DIF analyses based on CDM, similar approaches were performed, and

the reduced models were compared to a saturated model, G-DINA. Table 1 shows the results based on the relative fit indices obtained for the model selection that demonstrated the best fit to the data. Although the exact cutoff values for -2LL, AIC, BIC, CAIC, and SABIC relative fit indices have not been determined in the literature, the values of these indices used in model comparisons should be small.

Table 1

Comparing G-DINA to Reduced Models with Relative Fit Indices

Model	-2LL	AIC	BIC	CAIC	SABIC	χ^2	df	p-value
G-DINA	**14238.5**	**30848.5**	65649.1	73954.1	39289.3			
DINA	14983.7	31489.7	66072.4	74325.4	39877.7	745.2	52	<.001
DINO	15148.6	31654.6	66237.2	74490.2	40042.5	910.0	52	<.001
ACDM	14368.8	30918.8	65593.6	73868.6	39329.1	130.3	30	<.001
LLM	14304.7	30854.7	**65529.5**	**73804.5**	**39265.0**	66.17	30	<.001

G-DINA: *Generalized deterministic, noisy "and" gate,* ***DINA:*** *Deterministic, noisy "and" gate,* ***DINO:*** *Deterministic input, noisy "or" gate,* ***A-CDM:*** *additive CDM,* ***LLM:*** *linear logistic model.*

When the values of -2LL and AIC indices are examined, it is seen that G-DINA fits the data better than DINA, DINO, and ACDM. On the other hand, the BIC, CAIC, and SABIC indices show that the values in LLM are small, and the model fits the data better than G-DINA. The LR (likelihood ratio) test can be used to compare the more complex model (G-DINA) to the reduced model (LLM) in such situations (Ma & de la Torre, 2019b). The null hypothesis (H_0: The reduced model's fit to the data is as good as the more complex model) was tested in the LR test for this purpose, and the findings were reported in the table's "p-value" column. When referring to the table, it is clear that the LR test result is significant. G-DINA fits the data better than LLM, as demonstrated by as well. This study used MG G-DINA, a multi-group comparison extension of G-DINA, to provide diagnostic comparisons of male and female students' mathematics performance in the Wald test and LRT-based DIF analyses for the TIMSS 2011 assessment. Although MG G-DINA can be used for more than two groups, it was applied in this study for two different groups, as it did in Ma et al. (2021). All analyses for CTT, IRT, and CDM were performed in the R software using packages "GDINA" (Ma & de la Torre, 2020), "CDM" (Robitzsch et al., 2014), and "difR" (Magis et al., 2018). In the study, the p-values in determining the DIF for multiple comparisons between different methods were corrected using the Holm method, as Ma et al. (2021) used to control familywise error rates at the nominal level of .05.

Results

Results for CTT-based DIF Detection Methods

Findings were reported according to the MH and LR techniques, which are CTT-based DIF detection methods.

Figure 1 and Table 2 show the findings obtained using the MH technique.

Figure 1 displays DIF in seven items (X2, X3, X4, X14, X18, X20, and X27). Among these items, X18 moves far away from the critical value, while X27 moves slightly away from this value. This situation indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X27. Considering the p values in Table 2, and when the ΔMH values obtained from the MH technique for significant items are examined, it is seen that the findings in Figure 1 support the table, and seven items show DIF.

Figure 1

DIF Results Using the MH Technique

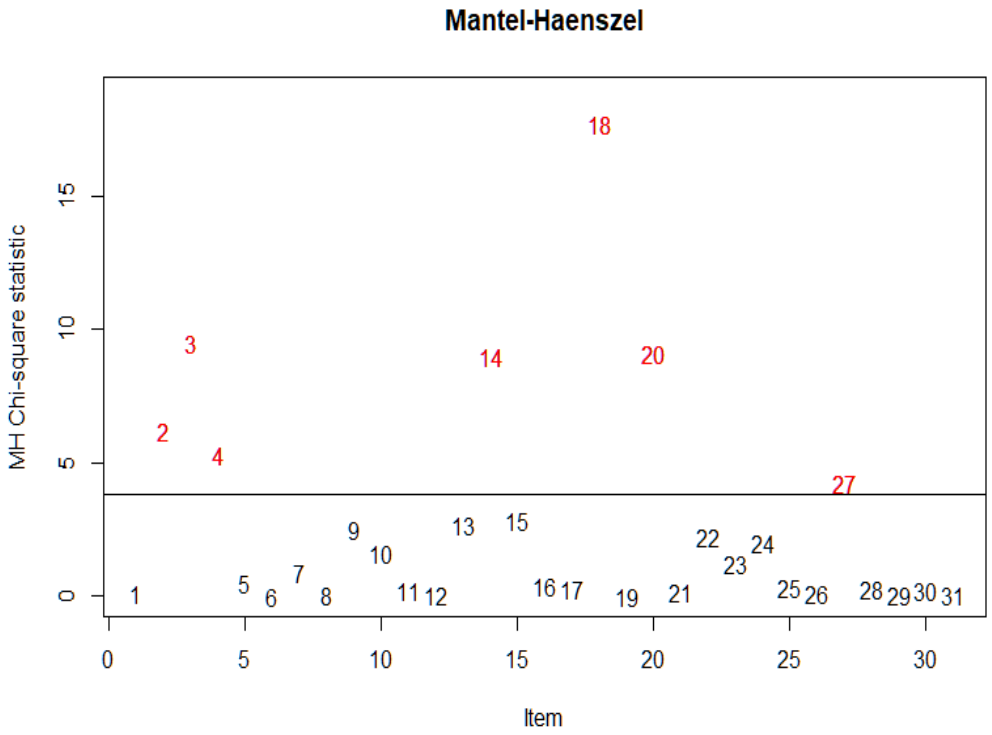


Table 2

DIF Results of the MH Technique

Item	χ^2	*p*	Alpha MH	ΔMH	Effect Size
X1	0.07	.78	0.92	0.19	A
X2	**6.16**	**.01***	0.58	1.26	B
X3	**9.50**	**.00***	2.33	-1.99	C
X4	**5.32**	**.02***	0.55	1.36	B
X5	0.50	.47	0.83	0.41	A
X6	0.00	.94	1.02	-0.05	A
X7	0.90	.34	0.77	0.61	A
X8	0.01	.90	1.14	-0.32	A
X9	2.49	.11	1.56	-1.04	B
X10	1.615	.20	1.41	-0.81	A
X11	0.19	.65	0.88	0.29	A
X12	0.04	.84	0.88	0.27	A
X13	2.66	.10	1.52	-0.99	A
X14	**9.01**	**.00***	0.39	2.17	C
X15	2.80	.09	1.53	-1.00	B
X16	0.35	.55	0.86	0.34	A
X17	0.28	.59	1.17	-0.37	A
X18	**17.71**	**.00***	3.29	-2.80	C
X19	0.00	.93	0.99	0.01	A
X20	**9.10**	**.00***	0.35	2.40	C
X21	0.13	.71	0.79	0.52	A
X22	2.23	.13	1.64	-1.17	B

Table 2

DIF Results of the MH Technique (Continued)

Item	χ^2	*p*	Alpha MH	ΔMH	Effect Size
X23	1.23	.26	0.74	0.69	A
X24	2.00	.15	0.67	0.91	A
X25	0.30	.58	0.85	0.36	A
X26	0.07	.78	1.12	-0.27	A
X27	**4.21**	**.04***	1.57	-1.06	B
X28	0.25	.61	0.88	0.28	A
X29	0.03	.84	0.93	0.16	A
X30	0.20	.65	1.13	-0.29	A
X31	0.02	.86	0.93	0.15	A

Effect size: 0=A; 1.0=B; 1.5=C *'A': Negligible effect; 'B': Moderate effect; 'C': Large Effect*

**p<.05*

Table 2 includes information on DIF's effect size and the magnitude of DIF. Four items (X3, X14, X18, and X20) have a large effect (C level) when the DIF levels of these items are evaluated. Three items (X2, X4, and X27) have a moderate effect (B level). ΔMH values have been examined to see if they were positive or negative, with "+" values favoring the focal group (female) and "-" values favoring the reference group (male). Items X2, X4, X14, and X20 provide an advantage for female students, whereas items X3, X18, and X27 provide an advantage for male students.

Figure 2 and Table 3 present the results obtained using the LR technique.

Figure 2

DIF Results Using the LR Technique

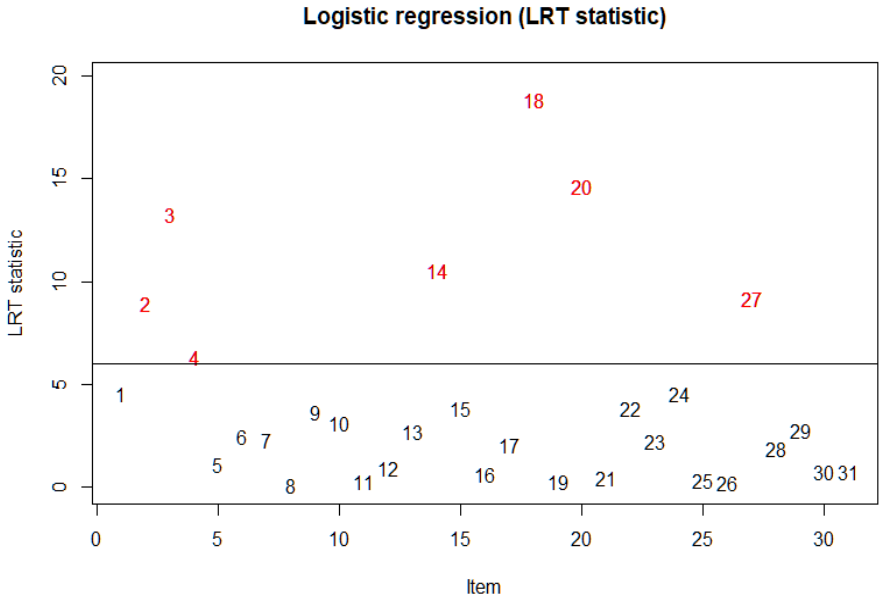


Figure 2 displays DIF in seven items (X2, X3, X4, X14, X18, X20 and X27). Among these items, it is seen that X18 moves far away from the critical value while X4 moves slightly away from it. This situation indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X4. The magnitude of the DIF was determined using the ΔR^2 values in Table 3.

Table 3

DIF Results of the LR Technique

Item	χ^2	*p*	ΔR^2	Effect Size
X1	4.56	.10	0.00	
X2	**8.92**	**.01***	0.06	A
X3	**13.24**	**.00***	0.02	A
X4	**6.33**	**.04***	0.01	A
X5	1.09	.57	0.00	
X6	2.48	.28	0.00	
X7	2.30	.31	0.00	
X8	0.08	.95	0.00	
X9	3.66	.16	0.00	
X10	3.13	.20	0.00	
X11	0.30	0.86	0.00	
X12	0.93	.62	0.00	
X13	2.72	.25	0.00	
X14	**10.56**	**.00***	0.03	A
X15	3.81	.14	0.00	
X16	0.63	.72	0.00	
X17	2.05	.35	0.00	
X18	**18.83**	**.00***	0.03	A
X19	0.28	.86	0.00	
X20	**14.60**	**.00***	0.02	A
X21	0.47	.79	0.00	
X22	3.85	.14	0.00	
X23	2.25	.32	0.00	
X24	4.53	.10	0.00	
X25	0.32	.84	0.00	
X26	0.19	.90	0.00	
X27	**9.16**	**.01***	0.02	A
X28	1.86	.39	0.00	
X29	2.77	.24	0.00	
X30	0.76	.68	0.00	
X31	0.72	.69	0.00	

Effect size: 0.01 = A; 0.13 = B; 0.26 = C

**p<.05,*

'A': Negligible effect; 'B': Moderate effect; 'C': Large effect

When Table 3 is examined, it is seen that seven items display DIF according to the LR technique. The DIF in these items is at the A level and has a negligible effect size, according to Zumbo and Thomas (1996)'s effect size (ΔR^2).

Results for IRT-based DIF Detection Methods

The findings were reported according to the Lord $\chi 2$ and Raju's Unsigned Area Measures Technique, which are IRT-based DIF detection methods, respectively. The findings obtained from the Lord $\chi 2$ technique are presented in Figure 3 and Table 4.

When Figure 3 is examined, it is seen that three red-colored items (X3, X18, and X20) above the threshold value show DIF. Among these items, it is seen that X18 moves far away from the critical value while X20 moves slightly away from it. This indicates that the largest DIF effect is in X18, and the lowest DIF effect is in X20.

Figure 3

DIF Results Using the Lord χ^2 Technique

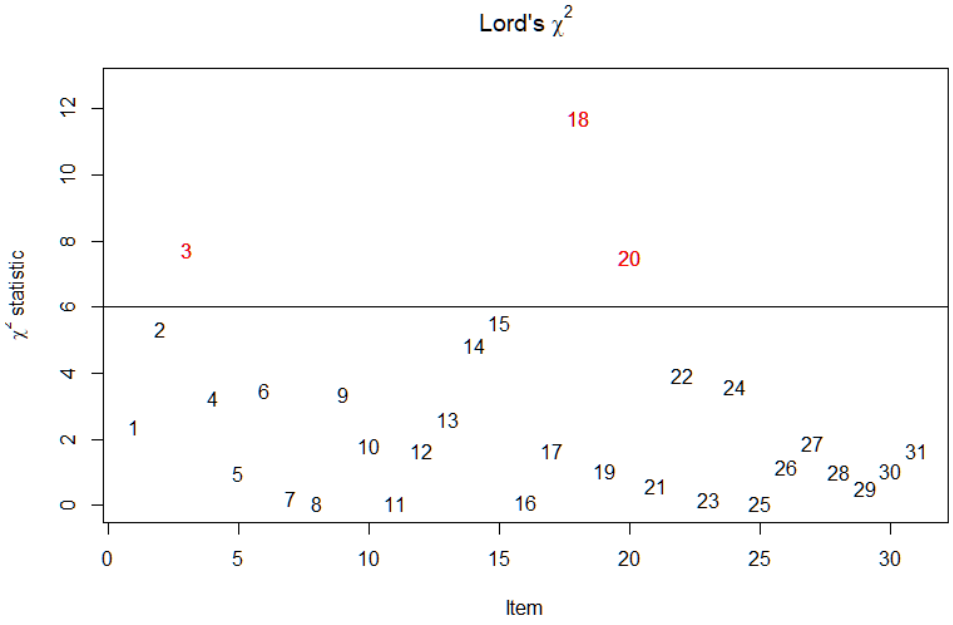


Table 4

DIF Results of the Lord χ^2 Technique

Item	Lord χ^2	*p*	Item	Lord χ^2	*p*
X1	2.38	.30	X17	1.65	.43
X2	5.34	.06	X18	**11.70**	**.00***
X3	**7.74**	**.02***	X19	1.04	.59
X4	3.26	.19	X20	**7.51**	**.02***
X5	0.96	.61	X21	0.60	.73
X6	3.48	.17	X22	3.91	.14
X7	0.22	.89	X23	0.17	.91
X8	0.05	.97	X24	3.58	.16
X9	3.35	.18	X25	0.06	.96
X10	1.80	.40	X26	1.15	.56
X11	0.05	.97	X27	1.88	.38
X12	1.64	.43	X28	1.00	.60
X13	2.59	.27	X29	0.53	.76
X14	4.83	.08	X30	1.04	.59
X15	5.53	.06	X31	1.64	.43
X16	0.10	.94			

**p<.05*

When Table 4 is examined, it is seen that three items show DIF according to the Lord χ^2 technique. Figure 4 and Table 5 display the results of Raju's unmarked area measures technique.

Figure 4

DIF Results Using the Raju's Unsigned Area Measures Technique

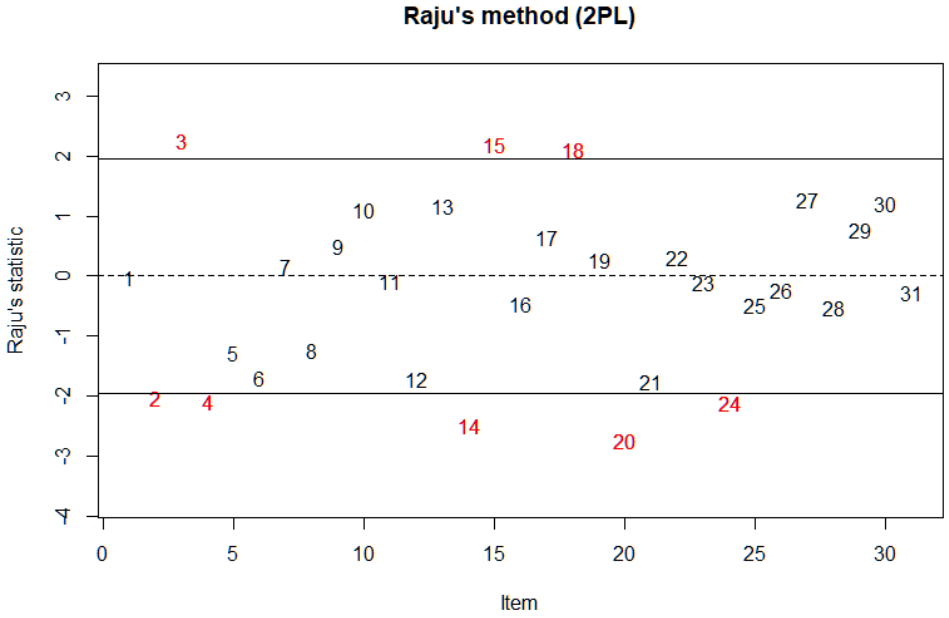


Figure 4 displays eight red-colored items (X2, X3, X4, X14, X15, X18, X20, and X24) above the threshold values that indicate DIF. Among these items, it is seen that X20 moves far away from the critical value while X2 moves slightly away from it. This indicates that the largest DIF effect is in X20, and the lowest DIF effect is in X2.

Table 5

DIF Results of the Raju's Unsigned Area Measures Technique

Item	Raju Statistic	*p*	Item	Raju Statistic	*p*
X1	-0.02	.97	X17	0.64	.52
X2	**-2.03**	**.04***	X18	**2.11**	**.03***
X3	**2.25**	**.02***	X19	0.27	.78
X4	**-2.09**	**.03***	X20	**-2.74**	**.00***
X5	-1.27	.20	X21	-1.75	.07
X6	-1.70	.08	X22	0.31	.75
X7	0.17	.86	X23	-0.10	.91
X8	-1.23	.21	X24	**-2.12**	**.03***
X9	0.49	.62	X25	-0.48	.62
X10	1.10	.26	X26	-0.23	.81
X11	-0.08	.93	X27	1.28	.19
X12	-1.71	.08	X28	-0.52	.60
X13	1.17	.23	X29	0.77	.43
X14	**-2.50**	**.01***	X30	1.21	.22
X15	**2.19**	**.02***	X31	-0.27	.78
X16	-0.46	.64			

**p<.05*

When Table 5 is examined, it is seen that eight items display DIF. Among these items, X2, X4, X14, X20, and X24 provide an advantage in favor of male students. It is seen that items X3, X15, and X18 provide an advantage in favor of female students.

Results for CDM-based DIF Detection Methods

This section used the Wald test and the LRT methods to determine if the test items indicate DIF, and the results are presented in Tables 6 and 7, respectively.

Table 6

DIF Results of The Wald Test

Item	Wald Statistic	Sd	*p*	*d-p*	DIF
X1	2.34	4	.67	1.00	-
X2	2.62	2	.26	1.00	-
X3	3.30	4	.50	1.00	-
X4	5.10	2	.07	1.00	-
X5	0.50	8	.00	1.00	-
X6	2.52	8	.96	1.00	-
X7	2.13	2	.34	1.00	-
X8	8.08	4	.08	1.00	-
X9	7.00	8	.53	1.00	-
X10	9.84	8	.27	1.00	-
X11	0.60	4	.96	1.00	-
X12	2.40	4	.66	1.00	-
X13	1.17	2	.55	1.00	-
X14	1.77	2	.41	1.00	-
X15	8.40	4	.07	1.00	-
X16	0.77	4	.94	1.00	-
X17	0.14	2	.93	1.00	-
X18	7.12	4	.12	1.00	-
X19	5.41	4	.24	1.00	-
X20	2.53	2	.28	1.00	-
X21	0.75	4	.94	1.00	-
X22	0.90	2	.63	1.00	-
X23	2.75	2	.25	1.00	-
X24	0.00	2	.00	1.00	-
X25	1.77	2	.41	1.00	-
X26	13.57	4	.00	.26	-
X27	7.86	4	.09	1.00	-
X28	0.51	2	.77	1.00	-
X29	0.72	2	.69	1.00	-
X30	3.98	4	.40	1.00	-
X31	**17.48**	**4**	**.00**	**.04**	+

'Sd': Degree of freedom; 'd-p': Adjusted p; '-': No DIF; '+': DIF

When the table is examined, it is clear that only one item (X31) displays DIF due to the Wald test. In the Q-matrix Sen and Arıcan (2015) utilized in their studies, this item was related to attributes 3 and 12. The findings obtained with LRT are presented in Table 7.

When the table is examined, it is seen that five items (X9, X10, X13, X20, and X30) show DIF with the LRT technique. Of these items, items X9 and X10 are associated with attributes 8, 9, and 10; item X13 is associated with attribute 13; item X20 is associated with attribute 4, and item X30 is associated with attributes 3 and 13 (Sen & Arıcan, 2015).

Table 7

DIF Results of LRT

	LRT Statistic	Sd	*p*	*d-p*	DIF
X1	4.95	4	.29	1.00	-
X2	5.52	2	.06	1.00	-
X3	9.29	4	.05	1.00	-
X4	-42.66	2	1,00	1.00	-
X5	0.85	8	.99	1.00	-
X6	12.89	8	.11	1.00	-
X7	3.62	2	.16	1.00	-
X8	-23.93	4	1.00	1.00	-
X9	**29.47**	**8**	**.00**	**.00**	+
X10	**53.90**	**8**	**.00**	**.00**	+
X11	-50.46	4	1.00	1.00	-
X12	9.42	4	.05	1.00	-
X13	**14.08**	**2**	**0.00**	**.02**	+
X14	4.59	2	.10	1.00	-
X15	-32.05	4	1.00	1.00	-
X16	11.02	4	.02	.65	-
X17	0.45	2	.79	1.00	-
X18	5.74	4	.21	1.00	-
X19	7.73	4	.10	1.00	-
X20	**33.25**	**2**	**.00**	**.00**	+
X21	10.10	4	.03	.92	-
X22	2.51	2	.28	1.00	-
X23	5.06	2	.08	1.00	-
X24	-0.00	2	1.00	1.00	-
X25	-7.60	2	1.00	1.00	-
X26	4.72	4	.31	1.00	-
X27	13.86	4	.08	.20	-
X28	1.21	2	.54	1.00	-
X29	-2.28	2	1.00	1.00	-
X30	**20.65**	**4**	**.00**	**.01**	+
X31	-23.93	4	.00	1.00	-

'Sd': Degree of freedom; 'd-p': Adjusted p; '-': No DIF; '+': DIF

The probability of having the attributes of interest and the prevalence according to the group's gender variable was investigated to better understand the items with DIF in CDMs and are given in Tables 8 and 9.

Students are assigned to one of the C latent classes using attribute probability. In the Turkey sample, there are 8,192 latent classes for 13 attributes. The prevalence estimate for an attribute is calculated by summing the probability for all relevant latent classes. Table 8 shows that the easiest attribute for male students is N10 (“Understands congruence and similarity using physical models, transparencies, or geometry software.”). About 58% of males have this attribute. The most difficult attributes for males are N1 (“Possesses an understanding of fraction equivalence and ordering; uses equivalent fractions as a strategy to add and subtract fractions”) and N5 (“Reasons about and solves one-variable equations and inequalities; uses properties of operations to generate equivalent expressions.”) because only 31% of males have these attributes.

Table 8

The Prevalence of Attribute by Gender

Attribute	Female	Male
N1	*.33*	*.31*
N2	.46	.40
N3	*.33*	.46
N4	.40	.38
N5	.38	*.31*
N6	.40	.32
N7	.44	.40
N8	.42	.38
N9	.34	.32
N10	.44	**.58**
N11	**.50**	.44
N12	.43	.33
N13	.41	.45

For female students, while the easiest attribute is N11 ("Recognizes perimeter, understands concepts of area, and relates area to multiplication and addition.") which is possessed by 50% of the students, the most difficult attributes are N1 and N3 ("Understands ratio concepts, and uses ratio reasoning to solve problems; finds a percent of a quantity as a rate per 100."), possessed by 33% of the students. In addition, it is seen that female students are more likely to master than male students in the remaining ten attributes except for N10, N3, and N13 ("Investigates chance processes and develops, uses, and evaluates probability models.").

Table 9

Profiles of Attributes of Students by Gender

Latent Class	Female	Male	Latent Class	**Female**	Male
0000000000000	**.14**	**.15**	0010000001000	.00	.04
0000000000100	.05	.04	0010000001100	.00	.04
0000000001000	**.07**	**.07**	0100100100000	.02	.00
0000000001100	.03	.00	0101001100111	.00	.02
0000010000000	.05	.00	1111101111111	.00	.02
0000110000000	.03	.00	1111111110111	.04	.00
0000110001000	.02	.00	**1111111111111**	**.07**	**.08**

When Table 9 is examined, it is seen that 14% of males and 15% of females are in the "0000000000000" latent class. That is, they have not mastered any of the attributes. In the latent class "11111111111111", which represents mastery of all attributes, 7% of females and 8% of males take place. In terms of comparison-based gender, although it is seen that males have a higher rate of mastering all attributes than females, the difference is about 1%. "0000000001000" is another common latent class. When this latent class is investigated, it is observed that only N10 is mastered by 7% of female and male students.

As seen in the findings obtained using MG G-DINA, it is seen that this model provides a diagnostic comparison of the mathematics performance of females and males in the TIMSS 2011 assessment within the scope of cognitive diagnostic assessments.

Comparison of the Methods

In this section, the responses given by male and female students who took the second booklet of the mathematics test in TIMSS 2011 Turkey sample to the items were analyzed to see if the items in the test displayed DIF or not and if the findings were presented in figures and tables according to gender.

The study results based on all methods are presented in Table 10, and comparisons of the methods are made.

Table 10

Comparison of DIF Results of Different Methods

	Traditional Methods					Based-CDM DIF Methods		
	CTT		*IRT*					
Item	*MH*	*LR*	*Lord* χ^2	*Raju*	**DIF**	*Wald*	*LRT*	**DIF**
X1	-	-	-	-	**0/4**	-	-	**0/2**
X2	+	+	-	+	**3/4**	-	-	**0/2**
X3	+	+	+	+	**4/4**	-	-	**0/2**
X4	+	+	-	-	**2/4**	-	-	**0/2**
X5	-	-	-	-	**0/4**	-	-	**0/2**
X6	-	-	-	-	**0/4**	-	-	**0/2**
X7	-	-	-	-	**0/4**	-	-	**0/2**
X8	-	-	-	-	**0/4**	-	-	**0/2**
X9	-	-	-	-	**0/4**	-	+	**1/2**
X10	-	-	-	-	**0/4**	-	+	**1/2**
X11	-	-	-	-	**0/4**	-	-	**0/2**
X12	-	-	-	-	**0/4**	-	-	**0/2**
X13	-	-	-	-	**0/4**	-	+	**1/2**
X14	+	+	-	+	**3/4**	-	-	**0/2**
X15	-	-	-	+	**1/4**	-	-	**0/2**
X16	-	-	-	-	**0/4**	-	-	**0/2**
X17	-	-	-	-	**0/4**	-	-	**0/2**
X18	+	+	+	+	**4/4**	-	-	**0/2**
X19	-	-	-	-	**0/4**	-	-	**0/2**
X20	+	+	+	+	**4/4**	-	+	**1/2**
X21	-	-	-	-	**0/4**	-	-	**0/2**
X22	-	-	-	+	**1/4**	-	-	**0/2**
X23	-	-	-	-	**0/4**	-	-	**0/2**
X24	-	-	-	-	**0/4**	-	-	**0/2**
X25	-	-	-	-	**0/4**	-	-	**0/2**
X26	-	-	-	-	**0/4**	-	-	**0/2**
X27	+	+	-	-	**2/4**	-	-	**0/2**
X28	-	-	-	-	**0/4**	-	-	**0/2**
X29	-	-	-	-	**0/4**	-	-	**0/2**
X30	-	-	-	-	**0/4**	-	+	**1/2**
X31	-	-	-	-	**0/5**	+	-	**1/2**

'-' No DIF; '+' DIF

As table 10 illustrates, when the MH and LR methods from CTT-based methods are compared, it is observed that both methods exhibit DIF for the same items (seven items). When the Lord's χ^2 (three items) and Raju's unsigned area measures (eight items) IRT approaches are compared, the X3, X18, and X20 items in both methods indicate DIF. Five more items indicate DIF using Raju's unmarked area measures technique. Besides, the technique marks the most DIF items among the six methods. Although there are differences between the traditional methods as a whole, it has been observed that X3, X18, and X20 items indicate DIF according to these traditional methods.

When CDM-based DIF detection methods are compared, the Wald test only indicates DIF in one item (X31), while the LRT indicates DIF in five items (X9, X10, X13, X20, and X30). In addition, three

items (X3, X18, and X20) that indicate DIF in conventional methods are investigated with CDM-based methods, and only item X20 indicates DIF with LRT. The items labelled as having DIF via LRT and Wald tests are totally different.

Discussion

Many psychometric questions regarding detecting DIFs in CDMs still exist. Investigating large-scale assessments in the context of DIF by adapting them into CDMs (Terzi & Sen, 2019) may be one of the disregarded questions because looking for meaning in an assessment without making inferences about validity would not give the expected benefit or have the desired influence on educational policies. The invariance of the parameters of the items in the TIMSS 2011 8th-grade mathematics test was controlled by comparing DIF determination methods based on CTT, IRT, and CDM to ensure the correct use of CDMs by performing a retrofitting study. The compatibility of the methods with each other was evaluated.

All assessments must be fair for students with different characteristics (ethnicity, social, and gender). Because DIF analyses are important in affecting groups' inferences from test items (Hou et al., 2014), the DIF effect has been determined using the gender variable as a variable for six different methods. As a result, the researchers' interest in DIF determinations in the test questions is expected to contribute to the validity of diagnostic assessments as an item with DIF can be a potential item for bias. For this purpose, MG G-DINA, which is one of the multi-group models used within the scope of cognitive diagnostic assessments and takes into account sample heterogeneity, was used. Within the scope of this model, this study was considered necessary in order to evaluate the performances of the Wald test and LRT methods, which are relatively newer than traditional methods, on real data.

Thirty-one items differed for both traditional methods and the methods within the scope of cognitive diagnostic assessments. When CTT-based MH and LR methods were compared, it was determined that the same items displayed DIF in both. When Lord's χ^2 and Raju's unsigned area measurements methods, both based on IRT, are compared, the items X3, X18, and X20 indicate DIF in both. DIF was also identified in five more items using Raju's technique of unsigned area measures. The findings show that CTT and IRT methods provide nearly identical outcomes in their own right. This situation supports the findings of previous studies (Kan et al.,2013; Odabas, 2016). Cokluk et al. (2016) stated that both CTT and IRT, produced with different methods on their own merits, are mostly consistent. When CDM-based DIF detection methods are compared in themselves, the Wald test detects DIF in only one item, whereas the LRT technique detects DIF in five. Furthermore, whereas the Raju Unmarked Area Measures technique in IRT had the largest DIF items, the Wald test technique developed for CDMs had the lowest DIF items. Only item X20 displays DIF with the LRT technique when the performances of three items that display DIF in traditional methods are examined using CDM-based methods. The items labelled as DIF by LRT and Wald tests are completely different. Odabas (2016), within the scope of his research, obtained a wide range of items labelled as DIF as a result of the analyzes performed under CDM under different conditions. So that comparisons should be made with the use of more than one technique for DIF studies in CDM.

In order to better comprehend items with DIF in CDMs, the prevalence and possibility of attributes were investigated in this study. The difference between the two groups is approximately 1% for the two most prevalent latent classes ("00000000000000", "11111111111111"), even though males exhibit greater rates of non-mastery and mastery of all attributes than females.

The findings indicated that methods not based on cognitive diagnosis models display DIF more than others. In contrast, the Wald test and LRT methods based on cognitive diagnosis models have fewer items with DIF. There may be several explanations for this situation. The first is that the test used was not developed within the scope of cognitive diagnostic assessments (Ravand & Baghei, 2019). Since determining the qualifications before the test development and defining the Q-matrix by developing the items related to these properties are the most important points of this evaluation approach, the test's psychometric properties may not have been fully determined due to the deficiencies experienced at this

point. However, as stated by the researchers, considering that the development and use of CDM-based tests are not easy and that the negative situations that may occur in ensuring the validity of the Q-matrix are taken into account, it is seen that many CDM applications are adapted to test data developed with non-CDM-based approaches in large-scale data (Gierl et al., 2010). Similar to this study, Odabas (2016) also developed and used a Q matrix prepared later for a previously developed exam in his research. In this process, the researcher stated that the interaction of matter and property in the Q matrix remained within certain limits. Despite this limitation, as stated by the researcher, it is thought that the preparation of the Q matrix and then the development of the exams will be effective in DIF studies as well as parameter estimation and classification accuracy within the scope of CDM. A second possible situation is that LRT may be sensitive to sample size in rejecting the hypothesis "There is no DIF in the relevant item." Mehrazmay et al. (2021) investigated the sensitivity of LRT to sample size and observed that the number of items with DIF increased when different sample sizes were examined. In their study, Ma et al. (2021) found that item discrimination had a significant impact on DIF determination and that Type I error rates in LRT increased when items had low discrimination. They also underlined that the Wald test tended to be conservative when the sample size was small and the item discrimination was high. Liu et al. (2019) reported that as the number of items with DIF increased, the power of MH and LRT methods decreased. Svetina et al. (2018) noted the difficulties with Q-matrix definitions affected the MH, Wald test, and LRT. These findings could explain inconsistencies in the methods utilized in terms of cognitive diagnostic assessments.

When evaluating the consistency of these methods, it is essential to remember that as the number of DIF items in the test increases, the meanings inferred from the scores decrease, raising questions about the validity of the results. As a result, additional research into these new methodologies is required, particularly in cognitive diagnostic assessments. It would be more effective to look into the contributions of these methods to the tests that have been developed, especially when considering the CDM-related test development processes. In this study, DIF analyses were performed, as in Milewski and Baron (2002), to determine the test's psychometric properties within the framework of CDMs rather than the source of bias. In addition, the results were compared with different DIF determination methods from traditional methods, and their compatibility was examined. DIF is not a direct indicator of bias. Due to the abilities of these subgroups, the items may have an actual effect. The source of the difference should be investigated before making a biased decision. Researchers can take the study further and make more comprehensive determinations about item bias in test structure, scope, and subgroups (Dorans & Holland, 1993) if they would like to. As a limitation of this study, the attribute structure of the DIF items was not examined. Future researchers should consider associating the structure of complexity of items with DIF. In addition, DIF analyses were based on only the gender variable. Researchers can also perform studies utilizing various variables.

Declarations

Author Contribution: Büşra EREN-Literature Review, Writing and Critical Review. Tuba GÜNDÜZ-Materials, Data Collection and Processing Analysis, Interpretation. Şeref TAN-Conception, Design and Supervision.

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

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

References

Akbay, L. (2021). Impact of retrofitting and item ordering on DIF. *Journal of Measurement and Evaluation in Education and Psychology*, *12*(2), 212-225. https://doi.org/10.21031/epod.886920

Asil, M., & Gelbal, S. (2012). Cross-cultural equivalence of the PISA student questionnaire. *Education and Science*, *37*(166), 236-249. https://eb.ted.org.tr/index.php/EB/article/view/1501

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

Cokluk, O., Gul, E., & Dogan-Gul, Ç. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences: Theory and Practice, 16*(1), 319-330. http://dx.doi.org/10.12738/estp.2016.1.0329

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model: development and applications. *Journal of Educational Measurement, 45*, 343–362. https://doi.org/10.1111/j.1745-3984.2008.00069.x

de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*(2), 253-273. https://doi.org/10.1007/s11336-015-9467-8

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa*, *20*(2), 89-97. https://doi.org/10 .1016/j.pse.2014.11.001

DiBello, L. V., & Stout, W. (2007). IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*, 285-291. https://doi.org/10.1111/j.1745-3984.2007.00039.x

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). Earlbaum. https://doi.org/10.1002/j.2333-8504.1992.tb01440.x

Gierl, M. J., Alves, C., & Majeau, R. T. (2010). Using the attribute hierarchy method to make diagnostic inferences about examinees' knowledge and skills in mathematics: An operational implementation of cognitive diagnostic assessment. *International Journal of Testing, 1*, 318-341. https://doi.org/10.1080/15305058.2010.509554

Hasancebi, B. (2021). *Farklı ölçek tiplerinde değişen madde fonksiyonunun belirlenmesi ve yöntemlerin karşılaştırılması üzerine bir çalışma* [A study on determination of item response function in different scale types and comparison of methods] *(Thesis No.687568)* [Doctoral dissertation, Karadeniz Teknik University]. Council of Higher Education Thesis. https://tez.yok.gov.tr/UlusalTezMerkezi/

Hou, L., de la Torre, J. D., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the wald test to investigate DIF in the DINA model. *Journal of Educational Measurement, 51*(1), 98-125. https://doi.org/10.1111/jedm.12036

Hou, L., Terzi R., & de la Torre, J. (2020). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education, 7*(2), 145-158. https://doi.org/10.21449/ijate.689752

Kan, A., Sünbül, Ö., & Ömür, S. (2013). Examination of the item functions of the 6th - 8th grade exams subtests according to various methods. *Mersin University Journal of the Faculty of Education, 9*(2), 207-222. https://dergipark.org.tr/tr/download/article-file/160893

Kang, C., Yang, Y., & Zeng, P. (2018). Q-Matrix refinement based on item fit statistic RMSEA. *Applied Psychological Measurement, 43*(527-542). https://doi.org/10.1177/0146621618813104

Lee, S., Han, S., & Choi, S. W. (2021). DIF detection with zero-inflation under the factor mixture modeling framework. *Educational and Psychological Measurement, 1*(21). https://doi.org/10.1177/00131644211028995

Li, F. (2008). *A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning* [Unpublished doctoral dissertation]. The University of Georgia.

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology, 10*, 11-37. https://doi.org/ 10.3389/fpsyg.2019.01137

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

Ma, W., & de la Torre, J. (2020). GDINA: An R package for cognitive diagnosis modeling. *Journal of Statistical Software, 93*, 1–26. https://doi.org/10.18637/jss.v093.i14

Ma, W., & de la Torre, J. (2019b). *GDINA: The generalized DINA model framework*. R package version (2.7.3). Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement, 45*(1), 37-53. https://doi.org/10.1177/0146621620965745

Magis, D., Beland, S., & Raiche, G. (2018). *difR: collection of methods to detect dichotomous differential item functioning (DIF)* (Version 5.0). https://CRAN.R-project.org/package=difR

Mantel, N. & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute, 22*, 719- 748. https://doi.org/10.1093/jnci/22.4.719

Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021) Detecting differential item functioning using cognitive diagnosis models: Applications of the wald test and likelihood ratio test in a university entrance

examination, *Applied Measurement in Education*, *34*(4), 262-284. https://doi.org/10.1080/08957347.2021.1987906

Milewski, G. B., & Baron, P. A. (2002, April, 2-4*). Extending DIF methods to inform aggregate reports on cognitive skills*. [Conference presentation]. The Annual Meeting of the National Council on Measurement in Education, New Orleans, LA. https://files.eric.ed.gov/fulltext/ED466712.pdf

Mullis, I. V.S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y. & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Odabas, M. (2016). *Değişen madde fonksiyonunu belirlemede DINA modelde işaretli alan indeksi, standardizasyon, ve lojistik regresyon tekniklerinin karşılaştırılması* [*The comparison of DINA model signed difference index, standardization and logistic regression techniques for detecting differential item functioning] (Thesis No.446894)* [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis. https://tez.yok.gov.tr/UlusalTezMerkezi/

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*, 267–281. https://doi.org/10.1177/0146621619858675

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://link.springer.com/article/10.1007/BF02294403

Ravand, H., & Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24-56. https://doi.org/10.1080/15305058.2019.1588278

Robitzsch, A., Kiefer, T., George, A. C., & Ünlü, A. (2014). *CDM: Cognitive Diagnosis Modeling* (Version 3.12). https://CRAN.R-project.org/package=difR

Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*.Guilford.

Sen, S., & Arıcan, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 238-253. https://doi.org/10.21031/epod.65266

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in Psychology*, *9*(696). https://doi.org/10.3389/fpsyg.2018.00696

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Terzi, R., & Sen, S. (2019). A nondiagnostic assessment for diagnostic purposes: Q-matrix validation and item based model fit evaluation for the TIMSS 2011 assessment. *SAGE Open*, 9, 1–11. https://doi.org/10.1177/2158244019832684

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* (Unpublished doctoral dissertation]. University of North Carolina at Greensboro.

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

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. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D., & Thomas, D. R. (1996, October). *A measure of DIF effect size using logistic regression procedures* [Conference presentation]. The National Board of Medical Examiners, Philadelphia. https://scholar.google.com/scholar?cluster=15614527111689986107&hl=tr&lr=lang_tr&as_sdt=2005&sciodt=0,5&as_ylo=20

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 95-105

The Impact of Missing Data on the Performances of DIF Detection Methods

Rabia AKCAN* Kübra ATALAY KABASAKAL**

Abstract

This study analyzed the impact of missing data techniques on performances of two differential item functioning (DIF) detection methods (Mantel Haenszel and Multiple Indicator and Multiple Causes) under missing completely at random missing data mechanism. Percentage of missing data was set at 5% and 15%. Zero imputation, listwise deletion and fractional hot-deck imputation were used to handle missing data. The data set of the study consisted of 17 items in the S12 item cluster of Programme for International Student Assessment (PISA) 2015 science test. Results showed that fractional hot-deck imputation produced the best results in identifying DIF items in all conditions and it had also the closest DIF values to the values obtained from complete data set. It was also found that multiple indicator and multiple causes method was more adversely affected than Mantel Haenszel by the presence of missing data.

Keywords: Differential item functioning, Mantel Haenszel, MIMIC, missing data.

Introduction

Missing data is a frequently encountered problem in quantitative research studies. Since standard statistical methods were designed for complete data sets, missing values create a significant problem for the researchers. Generally, researchers use various ad hoc methods to handle missing data before the analysis. An example of these strategies is discarding the cases with missing data (i.e., listwise deletion). Replacing missing values with variable mean is another method. Yet, these traditional methods can lead to significant bias in sample statistics (Peugh & Enders, 2004).

The rate of missing data, missing data mechanism and patterns of missing data should be considered in order to decide on the method to handle missing data. Rate of missing data is directly associated with the quality of statistical inferences. There is not a specified criterion in the literature with respect to a reasonable missing data rate to get valid statistical inferences (Dong & Peng, 2013). However, it is seen that the rate of missing data has mostly varied between 0% and 30% in previous studies (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Rousseau et al., 2004).

As previously stated, another aspect of handling missing data is to take the missing data mechanism into account. Rubin (1976) classified missing data mechanisms into three types: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Within the context of item responses, MCAR indicates that some examinees leave the item blank in a completely random way without a systematic mechanism related to the missingness. Data are MAR when the probability of an observation which includes missing data is directly connected with a measurable variable. The fact that male students' probability of leaving an item blank is higher than female students would be an illustration of MAR mechanism. MNAR mechanism refers to the case in which probability of being missing is related to the value of the variable itself. In this case, an examinee might leave the item unanswered as they do not know the answer (Finch, 2011b).

* Teacher, Ministry of National Education, Afyonkarahisar-Türkiye, eltrabia42@hotmail.com, ORCID ID: 0000-0003-3025-774X
** Assoc. Prof., Hacettepe University, Faculty of Education, Ankara-Türkiye, kkatalay@gmail.com, ORCIDID: 0000-0002-3580-5568

To cite this article:
Akcan, R., & Atalay-Kabasakal, K. (2023). The impact of missing data on the performances of DIF detection methods. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 95-105. https://doi.org/10.21031/epod.1183617

Received: 3.10.2022
Accepted: 22.03.2023

Missing data can affect quantitative research severely and may cause bias in parameter estimates, reduced statistical power, inflated standard errors and information loss (Dong & Peng, 2013). Therefore, it is essential for the researchers to investigate the impact of missing data on statistical techniques. Of particular concern in this research is the effect of missing data on the detection of differential item functioning (DIF), which causes systematic errors and reduces validity, with different methods. What follows is a brief overview of DIF and the methods used in this study.

DIF has received considerable attention as a result of the increased reliance on standardized achievement testing for evaluating the progress in education. The need to provide accurate assessment for all the examinees comes with great responsibility for psychometricians. Items intended to measure reading skills, for instance, must be suitable for the use with the students from various groups (e.g. gender, ethnicity etc.) to get meaningful score interpretations (Finch & French, 2007). If an item functions differently in a focal group compared with a reference group after controlling for differences in levels of performance on a latent trait (e.g., ability) of interest, it means the item shows DIF (Holland & Wainer, 1993; Scheuneman, 1979).DIF can be categorized into two broad types: Uniform and nonuniform. Uniform DIF is present when one of two groups has uniformly greater probability of answering an item correctly across all ability levels (Finch, 2005). Nonuniform DIF occurs when members of one group have greater probability in responding to an item correctly for some levels of the ability being measured, while they have lower probability for the other levels of the ability (Camilli & Shepard, 1994).

DIF detection methods can broadly be examined under two headings: (1) Classical Test Theory (CTT) and Item Response Theory (IRT). However, Camilli and Shepard (1994) highlighted that Confirmatory Factor Analysis (CFA) methods can be used to identify DIF as well. Previous studies in the field of DIF in the presence of missing data have mostly focused on CTT and IRT methods rather than CFA (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Rousseau et al., 2004). In this respect, we decided to use Mantel Haenzsel, a widely accepted method in literature based on CTT, and multiple indicator and multiple causes which is a CFA method becoming popular recently.

MIMIC

Multiple indicator and multiple causes (MIMIC) method is based on CFA and has received growing attention on DIF detection. The fundamental technique underlying DIF assessment with MIMIC models includes estimation of both direct and indirect effects for a grouping variable. The indirect effect shows whether there is a difference in the mean of latent trait across the groups, thereby explains the group differences on the latent trait. The direct effect shows whether response probabilities differ across the groups. In the DIF framework, MIMIC model can be written as (Finch, 2005):

$$y_i^* = \lambda_i \eta + \beta_i z_k + \varepsilon_i, \qquad (1)$$

where

y_i^* = latent response variable;

λ_i = factor loading for variable i;

η = latent trait;

β_i = slope relating the group variable with the response;

ε_i = random error; and

z_k = a dummy variable showing group membership.

Previous simulation studies investigating DIF with MIMIC method have shown that under most circumstances, MIMIC method performed as efficiently as or better than the other methods (SIBTEST, MH, LR etc.) with regard to type I error rate and power (e.g., Finch, 2005; Uğurlu & Atar, 2020; Woods, 2009). Missing data is a significant factor in the performances of statistical methods. Therefore, the impact of missing data on the DIF detection with MIMIC model is an important issue to be considered.

Mantel Haenszel

Mantel Haenszel (MH) statistic, proposed by Holland and Thayer (1988), might be the most commonly used among contingency table methods. With this method, probability of success on the item is compared for the members of two groups that are matched on the ability being measured. Firstly, respondents are divided into levels depending on the ability. Total test score is generally used for matching the respondents. For each score level, a 2x2 table is then created as in Table 1 (Clauser & Mazor, 1998).

Table 1

Data Organization in MH Method

Group	1 =Correct	0=Incorrect	Total
Reference	A_j	B_j	N_{Rj}
Focal	C_j	D_j	N_{Fj}
Total	M_{1j}	M_{0j}	T_j

MH statistic gives odds ratio (α), the ratio of the odds that reference group will respond to the studied item correctly to those for the focal group (Clauser & Mazor, 1998). Odds ratio is given in the equation (2).

$$\alpha = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j} \qquad (2)$$

Holland & Thayer (1988) recommended a logistic transformation to make interpretation of odds ratio easier. First, log of α is taken in order that the scale is symmetric around zero. Then, resulting value is multiplied by – 2.35 which produces Δ_{MH} (Clauser & Mazor, 1998).Zieky (1993) classified Δ_{MH} statistic into three categories: $|\Delta_{MH}| < 1$ shows negligible DIF (A level), $1 \leq |\Delta_{MH}| \leq 1.5$ shows moderate DIF (B level) and $|\Delta_{MH}| \geq 1.5$ shows large DIF (C level).

Returning briefly to missing data, it is obvious that presence of missing data is an important issue with regard to the DIF detection. However, commonly used DIF detection methods such as MH, SIBTEST and Logistic Regression (LR) are not capable of handling missing data. Hence, missing data handling methods used for the analysis might cause bias. Choice of missing data method may create DIF when there is no DIF in the item or eliminate DIF when it is actually present (Banks, 2015). When the choice of missing data handling method is inconvenient, erroneous decisions can be made based on DIF results which may prevent meaningful test score interpretations.

Researchers have attempted to assess the impact of missing data on DIF detection via simulation studies (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Garrett, 2009; Robitzsch & Rupp, 2009) or studies with real data (Rousseau et al., 2004; Tamcı, 2018). Most of these studies have focused on the widely used DIF detection methods such as SIBTEST, MH or LR. Emenogu et al. (2010) used both real and simulated data to investigate the impact of zero imputation (ZI), listwise deletion (LD) and analysis wise deletion on MH method. They reported that ZI produced false DIF regardless of the matching criterion used in the study and LD led to a significant decrease in sample size and the power of MH method.

Finch (2011b) also included IRT-LR in his study along with crossing SIBTEST and LR. This study has assessed the efficacy of ZI, LD, multiple imputation (MI) and stochastic regression imputation (SRI) on DIF detection. LD was recommended as a traditional missing data handling method for each DIF method and MI was the imputation method recommended in the study.

In recent years, there has been a growing amount of literature on the DIF detection with MIMIC, a CFA-based DIF detection method (Finch, 2005; Jin & Chen, 2020; Montoya & Jeon, 2020; Shih & Wang, 2009; Uğurlu & Atar, 2020; Woods, 2009). Missing data can affect any type of analysis including CFA (Harrington, 2009). Therefore, this study uses MIMIC method along with MH which is a broadly accepted method in the literature.

Zero imputation, listwise deletion and fractional hot-deck imputation (FHDI) were chosen as missing data handling method in the current study because the first two were widely used in prior research and far too little attention was paid to the last one. For ZI, all missing responses were replaced with 0. For LD, all individuals who had incomplete data responses were deleted. In FHDI, proposed by Kalton and Kish (1984) and investigated by Kim and Fuller (2004), M imputed values are created for each missing value, however, after fractional imputation a single data set is obtained as the output. Fractional weights are assigned to imputed values. The purpose of FHDI is to perform hot deck imputation efficiently (Im et al., 2015). FHDI was extended by Im et al. (2015) in two ways. First, in this new version of FHDI imputation cells are not required to be made in advance. Second, the proposed FHDI method is applied multivariate missing data with arbitrary missing patterns. In this paper, we used extension of FHDI proposed by Im et al. (2015) which is available in R software.

Purpose of the Study

DIF detection is an increasingly important area in test development and validity of standardized achievement tests which contribute to the development of educational policies (Zumbo, 2007). PISA (The Program for International Student Assessment), which enables comparison of students' achievement from different countries and languages and directs educational policies of these countries, is one of the important international standardized tests. Missing data can also be a problem in PISA application as with many other tests (e.g., Emenogu et al., 2010; Tamcı, 2018).

As already stated, traditional DIF detection methods cannot handle missing data. However, it is natural to have missing data in many educational or psychological tests. In this case, solving the missing data problem before DIF analysis becomes essential. Several studies investigating the missing data and DIF detection demonstrated that choice of missing data treatment method or type of missing data can have an influence on the DIF detection methods' performances (Finch, 2011a; Robitzsch & Rupp, 2009). This study therefore set out to assess the performances of DIF detection methods in PISA application in the presence of missing data. The leading research question in this investigation was as follows: What is the impact of (a) different missing data handling methods under (b) MCAR missing data mechanism and (c) different missing data percentages on the performances of the MH and MIMIC DIF detection methods?

Methods

This study aims to determine the impact of three missing data techniques on the performances of DIF detection methods under MCAR missing data mechanism. In this respect, this study is a descriptive study as it describes the existing situation as precisely as possible (Fraenkel et al., 2012).

Data Set

The data set consists of 17 items in the S12 item cluster of PISA 2015 science test. 1099 students from Finland who responded to all the items in the test were recruited as the sample of the study. Gender DIF studies are commonly carried out in international tests. However, gender DIF was not studied to make inferences on gender in this study. Different size of focal and reference groups might be another variable and affect the performances of missing data handling methods. As a result of this, Finland data set (1362 students) was chosen as the sample in the present study because the number of reference (female) and focal (male) groups was almost equal after discarding missing data.

Data Analysis

Data set includes 16 binary scored items and a partially scored item (CS637Q02S). This item was coded as 1-0 (full and partial point coded were as 1 and others were coded as 0) by the researchers and analyses were carried out on 17 items. A complete data set of 1,099 people (550 female and 549 male students) was obtained by discarding the missing data from the data set. After gender-based DIF analyses on complete data set were conducted with MH and MIMIC methods, results were recorded to be used as reference. Following DIF analyses, missing responses were created on complete data set by deleting data under MCAR. Missing responses under MCAR mechanism were created by selecting responses randomly from all items and all responses (0-1) for both groups. As the percentage of missing data mostly ranged between 0% and 30% in prior research (Banks & Walker, 2006; Finch, 2011a; Finch, 2011b; Robitzsch & Rupp, 2009; Tamcı, 2018), the percentage of missing data in the current research was set at 5% and 15%. Missing data were then dealt with ZI, LD and FHDI methods. DIF analyses were performed on these data sets. Finally, a comparison was made between reference DIF results and the results obtained from data sets that were completed with missing data handling methods. Whether numbers, levels or directions of DIF items in complete data set have changed or not was investigated. Pearson correlations of MH and MIMIC DIF statistics in all conditions were also examined. "MplusAutomation" (Hallquist & Wiley, 2018) and "difR" (Magis et al., 2010) packages were used for DIF analyses with MIMIC and MH methods respectively. Missing responses were generated in R through adapting the missing data codes written by Doğanay Erdoğan (2012). Imputation with FHDI method was conducted with "FHDI" (Im et al.,2018) package.

Results

Reference DIF results obtained from complete data set appear in Table 2. Those results were compared with DIF results of all combinations included in the study. We examined whether numbers, levels or directions of DIF items in complete data set have changed.

Table 2

DIF Results for MH and MIMIC Methods in Complete Data set

	MH		MIMIC
Item	Δ_{MH}	Level	Beta
Item1	0.071	-	0.038
Item2	-1.422	B (R)	-0.297*
Item3	-0.226	-	0.027
Item4	0.178	-	0.047
Item5	-0.815	-	-0.267*
Item6	-0.641	-	-0.110
Item7	0.491	-	0.089
Item8	0.332	-	0.153*
Item9	0.326	-	0.097
Item10	-1.313	B (R)	-0.218*
Item11	-0.750	A (R)	-0.178*
Item12	0.489	-	0.068
Item13	0.221	-	0.126
Item14	0.664	-	0.098
Item15	0.732	-	0.113
Item16	0.175	-	0.031
Item17	1.367	B (F)	0.220*

**Items showing DIF; R: Favors reference group F: Favors focal group*

As can be seen from the Table 2, two items displayed B level DIF and one item displayed A level DIF favoring reference group with MH method. One B level DIF item favoring focal group was also detected.

MIMIC method identified six DIF items. Four items (Item2, Item5, Item10 and Item11) favored reference group and two items (Item8 and Item17) favored focal group.

Table 3 illustrates DIF items with MH method in all combinations. DIF results were not reported for 15% missing condition with LD as it reduced sample size (70 students in total) dramatically. Sample size for 5% condition with LD was 457 students (232 students for the reference group and 225 students for the focal group).

Table 3

DIF Results for MH Under MCAR Mechanism

Method	5%	Δ_{MH}	15%	Δ_{MH}
Zero Imputation	Item2**(R)	-1.271	Item2*(R)	-0.698
	Item10*(R)	-0.963	Item8*(F)	0.712
	Item11*(R)	-0.774	Item11*(R)	-0.820
	Item17**(F)	1.054	Item17*(F)	0.912
Listwise Deletion	Item10***(R)	-1.863	--------	---------
	Item14**(F)	1.298		
	Item17***(F)	2.155		
Fractional Hot-Deck	Item2***(R)	-1.658	Item2**(R)	-1.344
	Item10***(R)	-1.591	Item10*(R)	-0.922
	Item11*(R)	-0.933	Item11**(R)	-1.158
	Item14**(F)	1.017	Item12*(F)	0.623
	Item17***(F)	1.500	Item15**(F)	1.210
			Item17**(F)	1.191

:Item showing A level DIF **:Item showing B level DIF *:Item showing C level DIF Significance level:0.05*

As shown in Table 3, directions of DIF items in complete data set did not change in all conditions. However, there have been differences in number of DIF items and DIF magnitude. Three missing data methods produced following results for 5% condition. DIF items remained the same with ZI, but DIF magnitude of one item (item10) decreased. When LD was used, two DIF items (item10 and item 17) did not change except that they had higher DIF value than their actual value. Item14 displayed DIF with LD although it was not among DIF items in complete data set. Four DIF items were identified correctly with FHDI, yet three of them were overestimated. Item14 showed false DIF in favor of focal group.

When the missing data percentage was 15%, ZI and FHDI both obtained false DIF. ZI identified three of the four DIF items in complete data set while FHDI identified them all. DIF magnitude of two items (item2 and item17) were underestimated with ZI. FHDI produced overestimated DIF magnitude for item11 while it underestimated the DIF magnitude of item10. Table 4 presents DIF items with MIMIC method in all combinations.

Table 4

DIF Results for MIMIC Under MCAR Mechanism

Method	5%	Beta	15%	Beta
Zero Imputation	Item2(R)	-0.278	Item8(F)	0.145
	Item5(R)	-0.256	Item11(R)	-0.195
	Item10(R)	-0.178	Item17(F)	0.153
	Item11(R)	-0.195		
	Item17(F)	0.157		

Table 4

DIF Results for MIMIC Under MCAR Mechanism (Continued)

Method	5%	Beta	15%	Beta
Listwise Deletion	Item9(F)	0.229		
	Item10(R)	-0.278	---------	--------
	Item17(F)	0.267		
	Item2(R)	-0.302	Item2(R)	-0.196
	Item5(R)	-0.293	Item8(F)	0.176
	Item8(F)	0.174	Item10(R)	-0.139
Fractional Hot-Deck	Item10(R)	-0.256	Item11(R)	-0.281
	Item11(R)	-0.185	Item13(F)	0.172
	Item13(F)	0.148	Item15(F)	0.200
	Item14(F)	0.197		
	Item17(F)	0.228		

**Items showing DIF; R: Favors reference group F: Favors focal group*

As in MH method directions of DIF items in complete data set did not change in all conditions for MIMIC method. Three missing data methods produced following results for 5% condition. ZI could not identify only one DIF item which had DIF in complete data set analysis. LD obtained false DIF for item9. Two out of six DIF items showing DIF in complete data set were determined as DIF items with LD.FHDI identified all DIF items accurately; however, it produced false DIF in favor of focal group in two items. In the case of 15% condition, both ZI and FHDI methods were unable to correctly identify all items indicating DIF in complete data set. Nevertheless, FHDI produced false DIF for this condition while ZI did not. Table 5 shows percentage of correctly identified DIF items and DIF free items by missing data handling methods with MH and MIMIC.

Table 5

Percentage of Correctly Identified DIF Items and DIF Free Items by Missing Data Handling Methods

DIF Detection Method	Missing Data method	5%	15%
MH			
	ZI	100%	75%
Percentage of correctly	LD	50%	----
identified DIF items	FHDI	100%	100%
Percentage of correctly	ZI	100%	92%
identified DIF free items	LD	92%	----
	FHDI	92%	85%
MIMIC			
	ZI	83%	50%
Percentage of correctly	LD	33%	----
identified DIF items	FHDI	100%	67%
Percentage of correctly	ZI	100%	100%
identified DIF free items	LD	90%	----
	FHDI	81%	81%

When examined in terms of the percentage of correctly identified DIF items and DIF free items in complete data set, it was found that for 5% condition with MH method, ZI and FHDI identified all DIF items in complete data set correctly. On the other hand, percentage of DIF items which were correctly identified by LD was 50%. DIF free items were the same with ZI. For this condition, 92% of DIF free items did not display DIF with LD and FHDI methods. FHDI determined all DIF items accurately for 15% missing case whereas percentage of DIF items obtained with ZI was 75%. Percentage of DIF free items which were correctly identified was 92% and 85% for ZI and FHDI methods respectively.

When MIMIC method was used, it was found that for 5% condition, FHDI identified all DIF items in complete data set accurately. Percentage of DIF items which were correctly identified were was 33% and 83% for LD and ZI respectively. Items that did not show DIF in complete data set were determined correctly with ZI. However, the percentage of DIF free items that were correctly identified by LD and FHDI were 90% and 81%.

FHDI was able to identify correctly 67% of DIF items for %15 missing case. The result was 50% for ZI in the same condition. ZI was better than FHDI in detecting DIF free items. ZI identified all DIF free items in complete data set correctly. On the other hand, the percentage of DIF free items correctly identified with FHDI was 81%. Table 6 provides correlations of MH and MIMIC DIF statistics in all conditions.

Table 6

Correlations of MH and MIMIC DIF Statistics in All Conditions

DIF Method	Complete Data	ZI (5%)	LD (5%)	FHDI (5%)	ZI (15%)	FHDI (15%)
MH						
Complete data	1	.975*	.826*	.985*	.825*	.913*
MIMIC						
Complete data	1	.968*	.671*	.981*	.795*	.808*

**Correlation is significant at the 0.01 level*

As Table 6 shows, all coefficients are positive and significant at p<.01. FHDI has the highest correlations for 5% and %15 missing case with both DIF methods. This result indicates that FHDI produces the closest DIF values to the values obtained from the complete data set. LD has the lowest correlation with both DIF methods. The correlations are slightly higher for MH method than MIMIC in all conditions.

Discussion

This study was designed to examine the impact of missing data techniques (ZI, LD and FHDI) on performances of MH and MIMIC DIF detection methods under MCAR missing data mechanism. Missing data percentage was set at 5% and 15%. The current study found that the percentage of identifying DIF items with LD was quite low for both DIF detection methods. It also produced the lowest correlations with reference DIF values regardless of the DIF detection method used. When the missing data percentage increased, sample size was reduced considerably with LD which resulted in no clear DIF results and could not be reported. This limitation was also reported by Emenogu et al. (2010) who could not calculate all DIF statistics with LD in their research.

Another important finding was that for both DIF detection methods, FHDI was the best in identifying the percentage of DIF items in all conditions while ZI was more successful than the other two methods in finding DIF free items. In terms of the correlations between the DIF statistics obtained from complete data set and the other conditions, FHDI had the highest correlations meaning it had the closest DIF values to the nonresponse data. ZI produced slightly lower correlations than FHDI. As regards to DIF detection methods, the results of the study indicated that the correlations are slightly higher for MH method than MIMIC in all conditions which suggests MIMIC method was more adversely affected than MH by the presence of missing data.

In the present study, the percentage of correctly identified DIF items with ZI was lower for the cases with higher missing data percentage regardless of the DIF detection method. Finch (2011b) reported that power rates for ZI decreased as the percentage of missing data increased in the study investigating the impact of missing data on nonuniform DIF detection. The most obvious finding of the current study was that LD was the least optimal method for both identifying DIF items and DIF free items in complete

data set. FHDI performed well in correctly identifying DIF items whereas ZI performed better than the other two methods in determining DIF free items in complete data set. In this case, the choice of missing data method should be based upon whether it is more essential to correctly identify items as DIF or falsely do so.

In this investigation, we aimed to study only with real data, which was a limitation of the research. There was only one sample size used in the study as we could not reach larger samples appropriate for our research. Relatively small sample size did not allow us to vary missing data rate; however, there might be missing data more than 15% in real life situations. Research is also needed to determine the performances of missing data handling methods (especially FHDI as it was the best of all) with larger samples and missing data rates.

As mentioned before there has been an increasing attention on DIF detection with MIMIC method. However, most studies in the literature have not dealt with DIF detection with CFA-based methods in detail when missing data is present. The aim of this study was to contribute to the literature on DIF detection with missing data by comparing two different methods based on CTT and CFA respectively. Since the study was limited to MH and MIMIC methods, it was not possible to see the performances of other methods based on CTT or IRT. Further work needs to be done to examine the performance of MIMIC method with missing data. Researchers might explore the effect of sample size, DIF magnitude and other missing data treatment methods on DIF detection with MIMIC and compare those results with DIF detection methods other than MH.

Declarations

Author Contribution: Rabia Akcan-Conceptualization, investigation, methodology, analysis, writing & editing. Kübra Atalay Kabasakal- Conceptualization, investigation, methodology, analysis, writing & editing, supervision.

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

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

References

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research & Evaluation*, 20(12), 1-10. https://eric.ed.gov/?id=EJ1059748

Banks, K., & Walker, C. (2006, April). Performance of SIBTEST when focal group examinees have missing data. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement Issues and Practice*, 17(1), 31-44. https://eric.ed.gov/?id=EJ564712

Doğanay Erdoğan, B. (2012). Çoklu atama yöntemlerinin Rasch modelleri için performansının benzetim çalışması ile incelenmesi [Assessing the performance of multiple imputation techniques for Rasch models with a simulation study] (Publication No. 314412) [Doctoral dissertation, Ankara University]. Council of Higher Education Thesis Center.

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *Springer Plus*, 2(1), 222. https://doi.org/10.1186/2193-1801-2-222

Emenogu, B. C., Falenchuk, O., & Childs, R. A. (2010). The effect of missing data treatment on Mantel-Haenszel DIF detection. *The Alberta Journal of Educational Research*, 56(4), 459-469. https://doi.org/10.11575/ajer.v56i4.55429

Finch, H. (2011a). The use of multiple imputation for missing data in uniform DIF analysis: Power and type I error rates. *Applied Measurement in Education*, 24(4), 281-301. https://doi.org/10.1080/08957347.2011.607054

Finch, H. (2011b). The impact of missing data on the detection of nonuniform differential item functioning. *Educational and Psychological Measurement*, 71(4), 663-683. https://doi.org/10.1177/0013164410385226

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, 29(4), 278-295. https://doi.org/10.1177/0146621605275728

Finch, H. W., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement*, 67(4), 565-582. https://doi.org/10.1177/0013164406296975

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* McGraw-hill.

Garrett, P. L. (2009). *A monte carlo study investigating missing data, differential item functioning, and effect size* (Publication No. 3401601) [Doctoral dissertation, Georgia State University]. ProQuest Dissertations Publishing.

Hallquist, M., & Wiley, J. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(4), 621-638. https://doi.org/10.1080/10705511.2017.1402334

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

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun, *Test Validity* (pp. 129-145). Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Im, J., Cho, I. H., & Kim, J. K. (2018). *FHDI: Fractional hot deck and fully efficient fractional imputation*. https://CRAN.R-project.org/package=FHDI

Im, J., Kim, J. K., & Fuller, W. A. (2015). Two-phase sampling approach to fractional hot deck imputation. *In Proceedings of the Survey Research Methods Section*, pages 1030-1043. http://www.asasrms.org/Proceedings/y2015/files/233957.pdf

Jin, KY., & Chen, HF. (2020). MIMIC approach to assessing differential item functioning with control of extreme response style. *Behavior Research Methods*, 52, 23-35. https://doi.org/10.3758/s13428-019-01198-1

Kalton, G., & Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics-Theory and Methods* ,13(16), 1919-1939. https://doi.org/10.1080/03610928408828805

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika*, 91(3), 559-578. https://doi.org/10.1093/biomet/91.3.559

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, 42(3), 847-862. https://doi.org/10.3758/BRM.42.3.847

Montoya, A. K., & Jeon, M. (2020). MIMIC models for uniform and nonuniform DIF as moderated mediation models. *Applied Psychological Measurement*, 44(2), 118-136. https://doi.org/10.1177/0146621619835496

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research:A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4), 525-556. https://journals.sagepub.com/doi/pdf/10.3102/00346543074004525

Robitzsch, A., & Rupp, A. A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, 69(1), 18-34. https://doi.org/10.1177/0013164408318756

Rousseau, M., Bertrand, R., & Boiteau, N. (2004, April). *Impact of missing data on robustness of DIF IRT-based and non IRT-based methods*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement*, 16, 143–152. https://www.jstor.org/stable/1433816

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, 33(3), 184-199. https://doi.org/10.1177/0146621608321758

Tamcı, P. (2018). Kayıp veriyle baş etme yöntemlerinin değişen madde fonksiyonu üzerindeki etkisinin incelenmesi [Investigation of the impact of techniques of handling missing data on differential item functioning] (Publication No. 517260) [Master's dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. *Journal of Measurement and Evaluation in Education and Psychology*, 11(1), 1-12. https://doi.org/10.21031/epod.531509

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44(1), 1-27. https://doi.org/10.1080/00273170802620121

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. Holland, & H. Wainer, *Differential Item Functioning* (pp. 337-347). Lawrence Erlbaum.

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