Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Kış* 2021 *Winter* 2021 *Cilt: 12- Sayı: 4* *Volume: 12- Issue: 4*

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 yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

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

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

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

Abdullah Faruk KILIÇ (Adıyaman Üni.)
Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Bolu Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Arife KART ARSLAN (Başkent Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengü BÖRKAN (Boğaziçi Üni.)
Betül ALATLI (Balıkesir Üni.)
Betül TEKEREK (Kahramanmaraş Sütçü İmam Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Ege Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Devrim ALICI (Mersin Üni.)
Devrim ERDEM (Niğde Ömer Halisdemir Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni.)
Duygu Gizem ERTOPRAK (Amasya Üni.)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Elif Kübra Demir (Ege Üni.)
Elif Özlem ARDIÇ (Trabzon Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Can AYBEK (Pamukkale Üni.)
Eren Halil ÖZBERK (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sütçü İmam Üni.)
Ersoy KARABAY (Kirşehir Ahi Evran Üni.)
Esin TEZBAŞARAN (İstanbul Üni.)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Ezgi MOR DİRLİK (Kastamonu Üni.)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni.)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Fuat ELKONCA (Muş Alparslan Üni.)
Fulya BARIŞ PEKMEZCİ (Bozok Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Görkem CEYHAN (Muş Alparslan Üni.)
Gözde SIRGANCI (Bozok Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni.)
Murat Doğan ŞAHİN (Anadolu Üni.)

Hakem Kurulu / Referee Board

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.)
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)
Sedat ŞEN (Harran Üni.)
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 (Aralık 2021, Sayı: 12-4)

Journal of Measurement and Evaluation in Education and Psychology (December 2021, Issue: 12-4)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 321-335

An Application of Latent Class Analysis for TIMSS 2015 Data: Detecting Heterogeneous Subgroups

Fatıma Münevver SAATÇİOĞLU *

Abstract
This study aimed to investigate the heterogeneity of the TIMSS 2015 data from Turkey and the USA 8th grade math. Latent Class Analysis (LCA) was used to determine the latent classes that cause heterogeneity in the data by using categorical observed variables. As a result of the LCA, supporting absolute and relative model fit indices through AvePP and entropy values, it was concluded that the data obtained from both countries fit the three-class model. The latent class probabilities and conditional response probabilities were examined for homogeneity and degree of segregation of the classes from each other. Based on the findings, it is recommended that the assumption of homogeneity in international evaluations be evaluated empirically with LCA. With this article, an example of the application of LCA is provided, and it is believed to be useful for researchers in the context of education and psychological evaluation.

Key Words: Latent class analysis, TIMSS 2015, heterogeneity.

INTRODUCTION

The correct understanding of study data is a significant factor for quality research in education and related fields. This is especially true for those investigating the role of scores in latent structures belonging to international large-scale assessment data. In large-scale international assessments such as Trends in International Mathematics and Science Study (TIMSS) and Programme for International Student Assessment (PISA), item and ability parameters are estimated using Item Response Theory (IRT) calibration. Despite many advantages, IRT models have strict assumptions such as unidimensionality, parameter invariance, local independence, and population homogeneity (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers,1991). In order to collect accurate evidence for the validity of the model, the assumptions of the model used in the analysis should be provided and there should be no biased items (Kreiner & Christensen, 2007). In some cases, the population may be heterogeneous due to the techniques or strategies used by individuals to correctly answer the items, familiarity with item content, etc (Embretson, 2007; Mislevy & Huang, 2007). In this case, it is not wise to use statistical models that require a single population in data analysis (Sen, 2016).

Different methods are used for the analysis of data obtained from heterogeneous populations. Analysis methods differ according to whether the population heterogeneity consists of observed or unobserved variables. If the variables causing heterogeneity are observed variables, some of the analysis methods used can be listed as discriminant analysis (DA), logistic regression (LR), multivariate analysis of variance (MANOVA), and multi-group factor analysis (MG-CFA). Of these methods, groups in LR, MANOVA, and MG-CFA are defined using a single observed variable or a combination of observed variables. DA and LR analyses are exploited if the goal is to identify variables to predict group membership, and MANOVA is preferred if it is aimed to compare group means by a set of observed variables. The MG-CFA, on the other hand, is designed for group comparisons by the means and covariances of a set of observed variables. Thus, the MG-CFA includes MANOVA as a submodel. In addition, these methods differ according to the type of observed outcome variables within a

* Lecturer.Dr., Rectorate, Yıldırım Beyazıt University, Ankara-Turkey, fmyigiter@gmail.com, ORCID ID: 0000-0003-4797-207X

To cite this article:
Saatcioglu F.M. (2021). An application of latent class analysis for TIMSS 2015 Data: Detecting Heterogeneous Subgroups. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 321-335. doi: 10.21031/epod.984771

Received: 19.08.2021
Accepted: 02.12.2021

subpopulation. If the observed variables are continuous, discriminant analysis (DA) and MANOVA are used, and LR analysis is employed if they are categorical. In MG-CFA, both categorical and continuous observed variables can be included in the same analysis (Lubke & Muthen, 2005).

Analyses using latent variables are person-centered techniques which are K-means clustering analysis, latent class analysis (LCA), and latent profile analysis (LPA) (Lazarsfeld & Henry, 1968; McLachlan & Peel, 2004; Magidson & Vermunt, 2002). Advantage of these person-centered techniques is that they provide a direct analytical translation of theories that hypothesize substantive and qualitative individual differences within the population. These methods are designed to detect clusters of participants with similar response patterns over a set of observed variables in a given dataset. The K-means method is based on an arbitrarily chosen criterion aiming to maximize inter-cluster variability while minimizing intra-cluster variability. LCA and LPA have additional advantages over cluster analysis: (a) individuals are assigned to latent classes based on conditional probabilities, and (b) models are statistically evaluated to decide the most appropriate model based on the observed data (Hagenaars & McCutcheon, 2002). Therefore, LCA and LPA appear in the literature as model-based methods in which alternative models are compared (Vermunt & Magidson, 2002). In LPA and LCA, a single categorical latent variable serves to model class membership (Lazarsfeld & Henry, 1968). For latent variables, analysis methods vary according to the type of observed variable. LPA is used if the latent variable is categorical and the observed variables are continuous, and LCA is favorable if the observed variables are categorical (Lubke & Muthen, 2005). LCA and LPA use multiple observed indicators (i.e., variables) to identify key population subgroups (i.e., latent classes) characterized by different behavioural patterns and are useful when it is not foreknown which participants belong to which subgroups (Butera, Lanza & Coffman, 2014). A latent categorical variable (i.e., underlying class membership) is used to model heterogeneity in the sample (Lubke & Muthén, 2005). In LCA and LPA, all covariation between observed variables is modelled to result from differences between classes. The observed variables within the class are independent of each other, which is called the local independence assumption (Goodman, 2002; McCutcheon, 2002). As it is the only assumption that needs to be met, LCA and LPA are flexible approaches that do not need many assumptions (Lubke & Muthen, 2005; Vermunt & Magidson, 2002). With the LCA, the profiles of the classes are determined by the classes obtained from students with similar reaction patterns (De Ayala & Santiago, 2017). Conditional item probabilities (probability of answering an item for students in a certain class) are used to label latent classes (Nylund, Asparouhov, & Muthén, 2007).

In recent years, latent class modelling techniques have attracted increasing attention among researchers due to the usability and developments in computer software for applications in the social and psychological sciences. Specifically, the use of LCA has increased in many areas such as health (Leech, McNaughton & Timperio, 2014; Olson, Hummer & Harris, 2017) and psychology (Chung, Park, & Lanza, 2005; Collins & Lanza, 2010; Lanza, Flaherty, & Collins, 2003). LCA helps to understand profile differences on multidimensional constructs (like personality, depression, etc.) and provides much more flexibility in parameterizing individual differences. Although LCA is used in various fields such as measurement invariance (Eid, Langeheine & Diener, 2003; Güngör, Korkmaz & Sazak, 2015; Güngör Çulha & Korkmaz, 2011; Kankaras, Moors & Vermunt, 2011; Morin, Meyer, Creusier, & Biétry, 2016; Yandı, Köse & Uysal, 2017), longitudinal latent growth models (Jung & Wickrama, 2008; Rindskopf, 2003) and Differential Item Functioning-DIF (Oliveri, Ercikan, Zumbo, & Lawless, 2014; Samuelsen, 2005; Uyar, 2015).

There are studies investigating latent classes using LCA in large-scale assessments (DeMars & Lau, 2011; Oliveri et al., 2014; Oliveri & von Davier, 2011; Rutkowski, 2018; Toker, 2016), but no exhaustive study has been found on how to apply it step-by-step in TIMSS data. In these studies, latent classes were determined with LCA, and it was indicated that the comments made would cause some adverse conditions due to the fact that IRT assumptions could not be met in the presence of latent classes. First, the presence of more than one latent class in the data obtained from the tests means that the measured structure changes for different classes, and this poses a threat to the validity of the test (Kreiner & Christensen, 2007; Messick 1994; Toker, 2016). It is because providing the assumptions of the model used in data analysis is regarded to be a requirement of ensuring the construct validity (Kreiner &

Christensen, 2007). Second, it is not fair to compare students with the same ability level as the assumptions of the IRT model are violated as a result of detecting latent classes in the data of large-scale assessments (Baghaei & Carstensen, 2013; Embretson, 2007; Oliveri & von Davier, 2011; Rutkowski & Rutkowski, 2018). Another problem is that the parameters estimated using the IRT model may be biased in the presence of different subgroups (DeMars & Lau, 2011; Park, Lee & King, 2016). Therefore, revealing the latent classes that cause heterogeneity in international large-scale test data is highly important in order to be able to analyze the test data correctly and to obtain accurate estimations. In addition, it is hoped that this study will contribute to the literature in terms of providing information on how to apply the LCA analysis to TIMSS 2015 data, how to test its assumptions and how to interpret the analysis outputs.

Purpose of the Study

This study aims to present an example of how to apply the LCA to TIMSS 2015 8th grade math data and to reveal the latent classes.

METHOD

In this study, latent class analysis was used based on students' responses to the items for TIMSS 2015 data. The data were analysed using the maximum likelihood estimation (MLE) method in the Mplus software program (Muthén & Muthén, 2017).

Sample

This study included 432 students from Turkey and 727 students from the USA. In this research, from these countries, the USA was chosen as the country with large sample size while Turkey was chosen as the country with medium sample size. Also, by 8th-grade math achievement average in TIMSS 2015, Turkey ranked 24th, and the USA ranked 10th among 39 OECD countries that took the exam (Mullis, Martin, Foy & Hooper, 2016). Accordingly, it can be alleged that Turkey has a medium level of achievement and the USA has a high level of achievement. So these two countries, which differ in sample size and success ranking, were selected.

Data Collection Tools

TIMSS 2015 is a standardized test that allows 4th and 8th-grade students of countries to determine their knowledge about concepts and processes in math and science, and their attitudes towards these subjects (Thomson, Wernert, O'Grady & Rodrigues, 2017). The instrument of the study is the math achievement test applied to 8th-grade students participating in TIMSS 2015. In the TIMSS assessment, items were developed in accordance with the cognitive processes of knowing, applying, and reasoning. About half of the items in the math test were multiple-choice while the other half consisted of long/short answered items. In TIMSS 2015 with science and math tests, the items in the achievement test included 28 blocks, 14 of which were science and 14 were math. The number of items in the booklets ranged from 11 to 17 (Martin, Mullis & Hooper, 2016). Since the 7th booklet contains more multiple-choice items (17 items), this booklet was chosen and analyzes were carried out. Eight of these items were for measuring knowing, six for applying, and three for cognitive reasoning domains.

Data Analysis

Latent class analysis is one of the finite mixture models used in social, behavioral, and health sciences to determine whether students are divided into latent classes based on a latent structure (Collins & Lanza, 2010). The purpose of LCA is to determine the class membership by using the categorical observed variables. LCA allows the analysis of dichotomously scored (1-0), ordinal and categorical variables, and

the combination of these variables (Nylund, Asparouhov & Muthén, 2007). Figure 1 illustrates below the relationship between latent and observed variables, with the *c* categorical latent variable (u1, u2, u3) for LCA representing the observed variables:

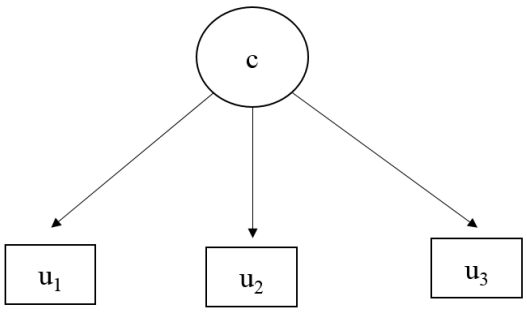

Figure 1. LCA Model for Latent and Observed Variables

The latent variable can be explained as unobservable variables determined by directly measured observed variables. The latent class, on the other hand, represents a statistically determined group of students with homogeneous response patterns, and different latent classes contain different homogeneous response patterns to items (Bolt, Cohen, & Wollack, 2001). In other words, it can be claimed that students in the same latent class have similar abilities, problem-solving skills, and answering strategies (Embretson, 2007; Glück & Spiel, 2007).

If $Y_{ij} \in \{0,1\}$ is the variable showing the responses of individual $i \in \{1,2,..,N\}$ to the items; $j \in \{1,2,\ldots,T\}$ and $g \in \{1,2,\ldots,G\}$ is the variable for the latent class membership of the individual, the probability of answering the items correctly by the individuals in a class $P(Y_{ij} = 1)$ can be equated as follows:

$$P(Y_{ij} = 1) = \sum_{g=1}^{G} \pi_g P(Y_{ij} = 1 | G = g) \quad (1)$$

In this equation, π_g represents the probability of the latent class and the conditional probability of $P(Y_{ij} = 1 | G = g)$ demonstrates the probability of answering item *j* correctly for the individual *i* in the g class.

As shown in Equation 1, the probability of obtaining an answer of Y_{ij} is the weighted average of the class-specific probabilities. The parameters to be estimated in the latent class model are the latent class probabilities and the conditional response probabilities (Nylund, Asparouhov & Muthén, 2007). These parameters help to examine the degree of homogeneity and latent class separation when evaluating model-data fit. The latent class probability parameters show the population ratio of the students in each latent class. The homogeneity of a latent class means that the students in the class have the same observed response pattern. The fact that the probability of responding to the variables observed in the latent class condition is 0 or 1 gives evidence that the latent classes are homogeneous. The conditional response probability parameters are interpreted while examining the separation of the latent classes. Latent classes are highly differentiated when the conditional response probabilities that are high in one latent class are low in another latent class.

MLE method is used to estimate the latent class analysis parameters. MLE is used to obtain parameter estimates by fitting a given latent class population model to the observed sample data. For mixture models, the likelihood function can generally be obtained by estimating full-information maximum likelihood (FIML) under the assumption of missing-at-random-MAR (McLachlan & Peel 2004). The MLE method continues the estimation of the parameter starting from the initial values until it finds the maximum probability of the parameter. When estimation is not started with appropriate initial values or there is a problem in defining the model, it can give the local maxima value instead of the global maximum of the estimated probability distribution. Estimating the model by taking different random

initial values with the STARTS and STITERATIONS commands added to the syntax in the Mplus software can provide a practical solution to this problem (Jung & Wickrama, 2008; Wang & Wang, 2012). If the problem cannot be dealt with despite taking these precautions, the source of the problem should be determined by examining the defined model, the number of latent classes, observed variables, and sample size.

The assumptions of latent class analysis

Although the latent class analysis does not have assumptions such as normal distribution and unidimensionality, it is necessary to provide the assumption of local independence, indicating that the observed variables are independent. This assumption means that the observed variables are interconnected only with the latent variables and there is no relationship between the errors of the observed variables (Vermunt & Magidson, 2004). To check this assumption, bivariate residuals for observed pairs of variables are examined. High scores of these values suggest local dependency (Vermunt & Magidson, 2004).

Model selection

When the number of classes in a population is unknown in advance, an exploratory approach is followed to determine the number of latent classes. This explanatory approach involves fitting models containing an increasing number of classes to the data and then finding the best fit among these candidate models. As a result of the analysis, the fit indices are compared and the model that best fits the data is selected (Sen, 2016). In determining the number of classes, many factors should be considered, including the research question, parsimony, theoretical justification, and interpretability as well as fit indices (Lubke & Neale, 2006). The principle of parsimony is choosing a model with fewer parameters instead of more complex ones. If latent classes are defined, each latent class must be significant and interpretable. Moreover, even if the model meets all the requirements of mathematical analysis, the predictive model will not be useful if it cannot provide a theoretically interpretable latent class (Wang & Wang, 2012). Therefore, fit indices and model fit tests should not be the decisive factors when deciding on the number of classes. In their simulation study, Nylund et al. (2007) determined that BLRT, followed by BIC, and then sample-size adjusted BIC (SSA-BIC) monitored the best performance among all fit indices and tests found in the Mplus output. However, Nylund et al. (2007) pointed out the disadvantages of the BLRT index due to the increased computation time of BLRT and its dependence on distributional and model assumptions. For example, if the data within a class is skewed but modelled as if the data were normally distributed, the BLRT *p*-value may be misinterpreted. Hence, Nylund et al. (2007) suggested interpreting the significance of the BIC and SSA-BIC value and the *p*-value obtained from the VLMR test as a guide to arrive at possible solutions in the first steps of the model research. Therefore, the model with a low BIC and SSA-BIC value and a *p*-value of less than 0.05 in the VLMR test should be selected (Jung & Wickrama, 2008). The likelihood ratio test (G^2), one of the absolute fit indices, gives information about whether a model fits the data well or not (Agresti, 1990). All of these indices are included in the Mplus output. An additional consideration in model selection is the size of the smallest class. While a four-class model may best fit the data, the researcher should be able to justify the addition of this class if this additional class consists of relatively fewer individuals (e.g. proportionally <1% and/or numerically n<25) (Lubke & Neale, 2006). In addition to all these indices, it is useful to examine the AvePP and entropy values.

The examination of average posterior probabilities and entropy values

When the possible number of classes is optimal, students are assigned to the latent classes in the latent class analysis. Based on a student's response patterns, the probability of latent class membership is measured by the probability of posterior class membership (Wang & Wang, 2012). For this reason, it is very important to examine the mean of posterior probabilities (Average Posterior Probabilities-AvePP) and entropy value related to classification. The posterior mean of probabilities (AvePP) provides

information about how well a particular model classifies students into their classes. Students' AvePP values greater than .70 indicates that the separation of students into classes is successful (Nagin, 2005). The entropy value shows the uncertainty in the classification. A single entropy value is generated for the entire analysis. The entropy value greater than .80 means that the classification uncertainty is low (Clark & Muthen, 2009).

RESULTS

First, it was examined whether the model fit indices and the local independence assumption were provided as a result of the LCA applied to Turkey and the USA data. Then, parameter estimations were examined, and latent class profiles were interpreted based on latent class probabilities and conditional response probabilities.

The Analysis of Model Fit Indices

In the analyses, 1, 2, 3 and 4-class models were tested, respectively, and the obtained model fit indices were presented in Table 1.

Table 1. Fit Indices of Models Tested for Data from Turkey

Fit indices	1-class model	2-class model	3-class model	4-class model
AIC	9004.712	8372.403	8267.709	8253.825
BIC	9073.835	8514.716	**8483.212**	8542.519
SSA-BIC	9019.887	8403.647	**8315.021**	8661.097
LR Chi-Square Test	1378.362	58.857	1613.448	193.184
LR Chi-Square *p*-value	1.0000	1.0000	1.0000	1.0000
VLMR Test	-	668.309	140.694	49.883
VLMR *p*-value	-	0.0000	0.0000	0.4093
BLRT Test	-	668.309	140.694	49.883
BLRT *p*-value	-	0.0000	0.0000	0.0200

*$p<.05$

According to Table 1, the LR Chi-Square test, which is an absolute fit index, has an insignificant *p*-value for data from Turkey, indicating that the model data fit is achieved. When the BIC and SSA-BIC values of the relative fit indices are examined, it is clear that the 3-class model is the one with the lowest values. The results of VLMR and BLRT can be found in the Mplus output under the Technical 11 and Technical 14 sections, respectively. Here, both VLMR and BLRT show a statistically significant difference between the 2-class and 3-class models. This result implies that the 3-class model provides a significant improvement in model fit compared to the 2-class model. In the next step, the 3-class model is compared with the 4-class model. However, it was observed that the *p* value was not significant when the *p*-value of the VLMR test was examined through testing the 4-class model by adding a class on top of the 3-class model with the suggestion of Nylund et al. (2007). This finding reveals that adding one more class to the 3-class model does not improve the model-data fit. According to these results, it was concluded that the 3-class model fits the data better. The model fit indices of the USA data were submitted in Table 2.

Table 2. Fit Indices of Models Tested for Data from the USA

Fit indices	1-class model	2-class model	3-class model	4-class model
AIC	14653.187	13964.440	13814.329	13788.241
BIC	14731.199	14125.053	**14057.542**	14114.055
SSA-BIC	14677.219	14013.917	**13887.251**	13889.608
LR Chi-Square Test	2517.045	2313.400	2204.668	2264.156
LR Chi-Square *p*-value	1.0000	1.0000	1.0000	1.0000
VLMR Test	-	724.747	186.111	62.088
VLMR *p*-value	-	0.0000	0.0001	0.3291
BLRT Test	-	724.747	186.111	62.088
BLRT *p*-value	-	0.0000	0.0000	0.0000

*$p<.05$

According to Table 2, the LR Chi-Square test, which is the absolute fit index, has an insignificant *p*-value for the USA data demonstrating that the model data fit is achieved. When the relative fit indices BIC and SSA-BIC are examined, it is observed that the model with the lowest value is the 3-class model. In addition, it was observed that all *p* values were significant at the level of $\alpha = .05$, except for the 4-class model when the *p* values of VLMR tests were examined with the suggestion of Nylund et al. (2007). Therefore, it can be alleged that the model-data fit did not improve as a result of testing the 4-class model by adding a class to the 3-class model. According to the results obtained from the USA data, it was concluded that the 3-class model fits the data better. As a result, latent classes were identified in the TIMSS 2015 8th grade math data from Turkey and the USA, and the heterogeneity in the data was revealed.

The Examination of the Local Independence Assumption

Bivariate residuals (BVR) were examined for observed pairs of variables in testing the local independence assumption, which means that the observed variables are independent in the latent class condition (Vermunt & Magidson, 2004). Higher BVR values (standardized z-score) indicate the presence of local dependency. This information, available in technical output 10 in Mplus software (Muthén & Muthén, 2017), was given in Table 3.

Table 3. Bivariate Residuals (BVR) Examined in 3-Class Model for Turkey and USA Data

	TURKEY				USA			
Item pairs	Category1 Category1	Category1 Category2	Category2 Category1	Category2 Category2	Category1 Category1	Category1 Category2	Category2 Category1	Category2 Category2
M1-M2	-0.355	0.430	0.330	-0.296	0.881	-0.876	-0.669	0.543
M1-M3	0.189	-0.407	-0.370	0.396	-0.451	0.769	0.233	-0.302
M1-M4	-0.187	0.286	0.014	0.008	-0.003	0.139	0.045	-0.104
M1-M5	-0.302	0.477	0.177	-0.210	-0.946	0.743	0.470	-0.596
M1-M6	0.663	-0.907	-0.682	0.546	0.378	-0.380	-0.300	0.242
M1-M7	0.850	-0.759	-0.706	0.687	0.774	-0.512	-0.364	0.255
M1-M8	-0.593	0.929	0.376	-0.476	-0.310	0.562	0.165	-0.223
M1-M9	-0.183	0.204	0.062	-0.041	-0.442	0.413	0.250	-0.267
M1-M10	0.141	-0.275	-0.100	0.123	0.251	-0.103	-0.054	-0.016
M1-M11	0.178	-0.390	-0.204	0.261	0.881	-0.876	-0.669	0.543
M1-M12	-0.027	0.011	-0.003	0.028	-0.451	0.769	0.233	-0.302
M1-M13	-0.377	0.573	0.312	-0.288	-0.003	0.139	0.045	-0.104
M1-M14	0.853	-0.839	-0.775	0.711	-0.946	0.743	0.470	-0.596
M1-M15	0.505	-0.529	-0.516	0.578	0.378	-0.380	-0.300	0.242
M1-M16	-0.451	0.886	0.369	-0.385	0.774	-0.512	-0.364	0.255
M1-M17	-0.472	0.638	0.327	-0.315	-0.310	0.562	0.165	-0.223

When Table 3 is examined, it is observed that the standardized residual values for all variable pairs are near 0. This finding shows that the observed variables in each latent class condition are independent of each other in 3-class model estimation. Accordingly, it is concluded that there will be no local dependency bias in the estimated parameters as there is no local dependency between the observed variables.

The Examination of the Estimated Parameters

The estimated latent class probability parameters and conditional response probability parameters for Turkey data are presented in Table 4. The probability parameters given in parentheses in the table for each latent class represent the population ratio in each latent class. Conditional response probabilities to observed variables under the latent class membership condition were given in Table 4.

Table 4. Parameter Estimations Obtained from 3-Class Model for Data from Turkey

	Levels	Class1 (0.13)	Class2 (0.35)	Class3 (0.52)
M1	0	0.000	0.026	0.588
	1	1.000	0.974	0.412
M2	0	0.130	0.455	0.660
	1	0.870	0.545	0.340
M3	0	0.000	0.514	0.835
	1	1.000	0.486	0.165
M4	0	0.660	0.884	0.787
	1	0.340	0.116	0.213
M5	0	0.471	0.846	0.781
	1	0.529	0.154	0.219
M6	0	0.000	0.111	0.679
	1	1.000	0.889	0.321
M7	0	0.142	0.444	0.620
	1	0.858	0.556	0.380
M8	0	0.419	0.879	0.760
	1	0.581	0.121	0.240
M9	0	0.190	0.506	0.653
	1	0.810	0.494	0.347
M10	0	0.083	0.532	0.784
	1	0.917	0.468	0.216
M11	0	0.356	0.657	0.835
	1	0.644	0.343	0.165
M12	0	0.146	0.812	0.856
	1	0.854	0.188	0.144
M13	0	0.196	0.425	0.771
	1	0.804	0.575	0.229
M14	0	0.006	0.026	0.310
	1	0.994	0.974	0.690
M15	0	0.044	0.097	0.509
	1	0.956	0.903	0.491
M16	0	0.184	0.604	0.854
	1	0.816	0.396	0.146
M17	0	0.246	0.526	0.706
	1	0.754	0.474	0.294

When Table 4 is examined, 13% of the students are in Class 1, 35% are in Class 2, and 52% are in Class 3. According to the conditional response probabilities, the students in Class 1 have a higher performance in answering the items correctly, the students in Class 2 have a moderate performance in answering the items correctly while the students in Class 3 have a lower level in answering the items correctly. The estimated latent class probability parameters and conditional response probability parameters for data from the USA were submitted in Table 5.

Table 5. Parameter Estimations Obtained from 3-Class Model for Data from the USA

	Levels	Class1 (0.18)	Class2 (0.56)	Class3 (0.26)
M1	0	0.021	0.079	0.344
	1	0.979	0.921	0.656
M2	0	0.020	0.319	0.616
	1	0.980	0.681	0.384
M3	0	0.278	0.594	0.719
	1	0.722	0.406	0.281
M4	0	0.091	0.401	0.727
	1	0.909	0.599	0.273
M5	0	0.283	0.842	0.726
	1	0.717	0.158	0.274
M6	0	0.280	0.554	0.744
	1	0.720	0.446	0.256
M7	0	0.069	0.286	0.442
	1	0.931	0.714	0.558
M8	0	0.141	0.450	0.775
	1	0.859	0.550	0.225
M9	0	0.077	0.151	0.403
	1	0.923	0.849	0.597
M10	0	0.140	0.222	0.523
	1	0.860	0.778	0.477
M11	0	0.063	0.134	0.665
	1	0.937	0.866	0.335
M12	0	0.345	0.812	0.876
	1	0.655	0.188	0.124
M13	0	0.010	0.180	0.514
	1	0.990	0.820	0.486
M14	0	0.020	0.203	0.713
	1	0.980	0.797	0.287
M15	0	0.315	0.690	0.698
	1	0.685	0.310	0.302
M16	0	0.652	0.859	0.843
	1	0.348	0.141	0.157
M17	0	0.148	0.561	0.693
	1	0.852	0.439	0.307

When Table 5 is examined, the latent class probability parameters given in parentheses represent the population ratio in each latent class. In other words, 18% of the students are in Class 1, 56% are in Class 2, and 26% are in Class 3. According to the conditional response probabilities, for example, students in latent Class 1 have a higher performance in answering the items correctly, students in Class 2 have a moderate performance in answering the items correctly, while students in Class 3 have a lower performance in answering the items correctly.

The Interpretation of Latent Class Profiles

When the three homogeneous classes obtained from Turkey and the USA data were examined, Class 1, which had a high probability of correctly responding to the items, was interpreted as a *high-performing class*. Class 2, in which the probability of answering the items correctly is moderate, has the characteristics of a *medium-performing class*. Class 3, where the probability of giving correct answers to the relevant items is low, was called the *low-performing class*.

The Examination of Classification Ratios

AvePP and entropy values were calculated to determine the practical usefulness of the model. These values obtained for data from Turkey and the USA were summarized in Table 6.

Table 6. Classification Rates for the 3-Class Model Obtained from Turkey and the USA Data

	Turkey		Class 1	Class 2	Class 3
Entropy	0.888	Class 1	0.912	0.088	0.000
		Class 2	0.034	0.867	0.099
		Class 3	0.000	0.079	0.921

	The USA		Class 1	Class 2	Class 3
Entropy	0.810	Class 1	0.889	0.110	0.001
		Class 2	0.050	0.866	0.084
		Class 3	0.000	0.139	0.861

When Table 6 is examined, it is observed that the entropy value, which gives an overall value of classification accuracy, is approximately 0.89 for Turkey data and 0.81 for the USA data. It can be asserted that the three-class model is useful in assigning students to the correct classes as the entropy values obtained for both countries are greater than .80 (Clark, 2010). Upon examining the AvePP values, which demonstrate the average of the class probabilities of the students with the maximum posterior probability, it is obvious that these values are above 0.86 for each latent class. These indicate that students have high maximum posterior probability values in being assigned to classes.

DISCUSSION and CONCLUSION

This study presented an example of how to apply the LCA in TIMSS 2015 data, and it was investigated whether the datasets were homogeneous through the LCA. As a result of the LCA, supporting absolute and relative model fit indices through AvePP and entropy values, it was concluded that the data obtained from both countries fit the three-class model. When bivariate residuals (BVR) were examined, it was seen that the observed variables in each latent class condition are independent of each other in 3-class model estimation. The latent class probabilities and conditional response probabilities were reported for homogeneity and degree of segregation of the classes from each other. As a result, the students in Class 1 have a higher performance in answering the items correctly, the students in Class 2 have a moderate performance in answering the items correctly, while the students in Class 3 have a lower level in answering the items correctly. For Turkey, 13% of the students are in Class 1, 35% are in Class 2, and 52% are in Class 3. Also, for the USA 18% of the students are in Class 1, 56% are in Class 2, and 26% are in Class 3. It can be seen that the percentage of the better performing class for American students is more than for Turkish students, while the percentage of the underperforming class for American students is less than for Turkish students. Toker (2016), in his research, examined four countries (Turkey, USA, Finland, Singapore) with different educational systems for TIMSS 2011 8th grade math data. Three latent classes were identified using the latent class analysis. Oliveri et al. (2014) addressed 4th students from Taiwan, Hong Kong, Qatar, and Kuwait who participated in PIRLS 2006. In order to reveal the heterogeneity in the response patterns of the students, three latent classes were determined through the latent class analysis approach. Based on the findings, it is recommended that the assumption of homogeneity in international evaluations be evaluated empirically with LCA.

Also, the indices used during the model determination process in this study, as a result of applying LCA to TIMSS 2015 data, support the simulation results performed by Nylund et al. (2007). It was determined that the BIC and SSABIC values obtained for the number of classes that best fit the model were low. In addition to the model fit indices, the number of classes was decided by examining the *p* significance value with the VLMR test.

In mixture models such as latent class analysis, the inclusion of auxiliary variables such as covariant variables in the analysis provides valuable information in understanding the population heterogeneity embodied by a latent class variable. In particular, with this approach, it can be determined whether there

are direct effects from covariates to latent variable indicators in an attempt to identify possible sources of DIF (Masyn, 2017).

Future work should focus on extending the classification to include other demographic variables such as gender, age, and socioeconomic status. In addition, it can be used to examine whether latent classes obtained from various distal outcomes (e.g., academic performance, self-efficacy, etc.) show statistically significant mean-level differences or whether these procedures can be included in the latent class determination procedure. Such studies can be used to increase the predictive and discriminant validity of the test. Therefore, they can contribute to test validity.

Latent profile analysis can be used to investigate students' attitude profiles and how these profiles are associated with academic achievement in a standard math and science test for the variables measured by graded Likert-type questionnaires in TIMSS and PISA test data (e.g., attitude). Defined profiles can be a useful way for math and science teachers to understand better the different types of students in their classrooms. Arrangements can be made in the education programs for the deficiencies of the students in the classes.

REFERENCES

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

Baghaei, P., & Carstensen, C. H. (2013). Fitting the mixture Rasch model to a reading comprehension test: Identifying reader types. *Practical Assessment, Rese*, *18*(5), 1–13. doi: 10.7275/n191-pt86

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*, 331-348. doi: 10.1111/j.1745-3984.2002.tb01146.x

Butera, N. M., Lanza, S. T., & Coffman, D. L. (2014). A framework for estimating causal effects in latent class analysis: Is there a causal link between early sex and subsequent profiles of delinquency? *Prevention Science*, *15*(3), 397– 407. doi: 10.1007/s11121-013-0417-3

Chung, H., Park, Y., & Lanza, S. T. (2005). Latent transition analysis with covariates: Pubertal timing and substance use behaviours in adolescent females. *Statistics in Medicine*, *24*, 2895-2910. doi: 10.1002/sim.2148

Clark, S. L. (2010). *Mixture modeling with behavioral data* (Unpublished doctoral dissertation). University of California, Los Angeles.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology*, *34*(2), 195-210. doi: 10.1177/0022022102250427

Clark, S. L., & Muthén, B. O. (2009). Relating latent class analysis results to variables not included in the analysis. 2009 Manuscript submitted for publication. Retrieved from http://www.statmodel.com/download/relatinglca.pdf

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New Jersey: Wiley.

De Ayala, R. J., & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, *60*, 25-40. doi: 10.1016/j.jsp.2016.01.002

DeMars, C. E., & Lau, A. (2011). Differential item functioning detection with latent classes: How accurately detect who is responding differentially? *Educational and Psychological Measurement*, *71(4)*, 597–616. doi:10.1177/0013164411404221

Embretson, S. E. (2007). *Mixture Rasch models for measurement in cognitive psychology*. In M. von Davier & C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 235-253). New York: Springer Verlag.

Glück, J., & Spiel, C. (2007). *Studying development via Item Response Model: A wide range of potential uses*. In M. von Davier, & C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 281-292). New York: Springer Verlag.

Goodman, L. A. (2002). *Latent class analysis: The empirical study of latent types, latent variables, and latent structures*. In J. A. Hagenaars & A. L. McCutcheon (Eds.), Applied latent class analysis. New York: Cambridge University Press.

Güngör Culha, D. & Korkmaz, M. (2011). Örtük sınıf analizi ile bir örnek uygulama. *Eğitimde ve Psikolojide Ölçme Değerlendirme Dergisi*, *2*(2), 191-199.

Güngör, Culha, D., Korkmaz, M. & Somer, O. (2013). Çoklu-grup örtük sınıf analizi ve ölçme eşdeğerliği. *Türk Psikoloji Dergisi*, *28*(72), 48-57.

Hagenaars, J. A., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge, UK: Cambridge University Press.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Jeon, M. (2018). A constrained confirmatory mixture IRT model: Extensions and estimation of the Saltus model using Mplus. *The Quantitative Methods for Psychology*, *14*, 120–136. doi: 10.20982/tqmp.14.2.p120

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. doi: 10.1111/j.1751-9004.2007.00054.x

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, Item Response Theory, and latent class approaches. *Sociological Methods & Research*, *40*, 279–310. doi: 10.1177/0049124111405301

Kreiner, S., & Christensen, K. B. (2007). *Validity and objectivity in health-related scales: Analysis by graphical loglinear Rasch models*. In M. von Davier & C. H. Carstensen (Eds.). Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 329-346). New York: Springer Verlag. doi:10.1007/s11336-013-9347-z

Lanza, S. T., Flaherty, B. P., & Collins, L. M. (2003). *Latent class and latent transition analysis*. In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology: Vol. 2, research methods in psychology (pp. 663-685). Hoboken, NJ: Wiley.

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

Leech, R. M., McNaughton, S. A., & Timperio, A. (2014). The clustering of diet, physical activity and sedentary behavior in children and adolescents: A review. *International Journal of Behavioral Nutrition and Physical Activity*, *11*, 4. doi: 10.1186/1479-5868-11-4

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767-778. doi: 10.1093/biomet/88.3.767

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

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, *41*, 499–532. doi: 10.1207/s15327906mbr4104_4

Magidson, J., & Vermunt, J. K. (2002). Latent class models for clustering: A comparison with K-means. *Canadian Journal of Marketing*, *20*(1), 36–43.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. Zugriff am (Vol. 21).

Masyn, K. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling*, *24*, 180-197. doi: 10.1080/10705511.2016.1254049

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. New York: Wiley.

Morin, A. J., Meyer, J. P., Creusier, J., & Biétry, F. (2016). Multiple-group analysis of similarity in latent profile solutions. *Organizational Research Methods*, *19*(2), 231–254. doi: 10.1177/1094428115621148

Mullis, I. V., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Amsterdam, The Netherlands: International Association for the Evaluation of Educational Achievement.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23. doi: 10.3102/0013189X023002013

Mislevy, R., & Huang, C. W. (2007). *Measurement models as narrative structures*. In M. von Davier & C. H. Carstensen (Eds.), Multivariate and mixture distribution Rasch models: Extensions and applications. New York: Springer Verlag.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th Edition). Los Angeles, CA: Muthén & Muthén.

Nagin, D. (2005). *Group-based modeling of development*. London: Harvard University.

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

Oliveri, M. E., Ercikan, K., Zumbo, B. D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments—Comparing a latent class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265-287. doi: 10.1080/15305058.2014.891223

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*(3), 315–333.

Olson J. S., Hummer A. K., & Harris K. M. (2017). Gender and health behavior clustering among U.S. *Young Adults, Biodemography and Social Biology*, *63*(1), 3-20. doi: 10.1080/19485565.2016.1262238

Park, Y. S., Lee, Y.-S., & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*, *7*, 255. doi: 10.3389/fpsyg.2016.00255

Rindskopf, D. (2003). Mixture or homogeneous? Comment on Bauer and Curran (2003). *Psychological Methods*, *8*(3), 364–368. doi: 10.1037/1082-989X.8.3.364

Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research*, *62*(3), 354–367. doi: 10.1080/00313831.2016.1261044

Samuelsen, K. (2005). *Examining differential item functioning from a latent class perspective*. Doctoral Dissertation. Available from ProQuest Dissertations and Theses database. (UMI No. 3175148).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annuals of Statistics*, *6*(2), 461-464.

Sen, S. (2016) Applying the Mixture Rasch Model to the Runco Ideational Behavior Scale. *Creativity Research Journal*, *28*(4), 426-434. doi: 10.1080/10400419.2016.1229985

Thomson, S., Wernert, N., O'Grady, E., & Rodrigues, S. (2017). *TIMSS 2015: Reporting Australia's results*. Melbourne, Australia: Australian Council for Educational Research.

Toker, T. (2016). *A comparison latent class analysis and the mixture Rasch model: A cross-cultural comparison of 8th grade mathematics achievement in the fourth international mathematics and science study (TIMSS-2011)*. Doctoral Dissertation, The Faculty of the Morgridge College of Education University of Denver, USA.

Uyar, Ş. (2015). *Gözlenen gruplara ve örtük sınıflara göre belirlenen değişen madde fonksiyonunun karşılaştırılması*. Yayınlanmamış Doktora Tezi. Hacettepe Üniversitesi, Eğitim Bilimleri Ensitüsü. Ankara.

Vermunt, J. K., & Magidson, J. (2004). *Latent class analysis*. The Sage Encyclopedia of Social Sciences Research Methods (pp. 549-553).

Vermunt, J. K., & Magidson, J. (2020). How to perform three-step latent class analysis in the presence of measurement non-invariance or differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*. doi: 10.1080/10705511.2020.1818084

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Hoboken, NJ: John Wiley & Sons.

Yandı, A., Köse, İ. A. & Uysal, Ö. (2017). Farklı yöntemlerle ölçme değişmezliğinin incelenmesi: PISA 2012 örneği. *Mersin Üniversitesi Eğitim Fakültesi Dergisi*, *13*(1), 243-253. doi: 10.17860/mersinefd.305952

Appendix. Mplus Code for LCA

```
TITLE:Booklet 7 2 Class Solution Latent Class Analysis
DATA:FILE IS data7.txt;
VARIABLE:     NAMES ARE M1-M17;
USEVARIABLES = M1-M17;
CATEGORICAL = M1-M17;
CLASSES = c (2);
MISSING ARE ALL (99);
ANALYSIS:TYPE = MIXTURE;
OUTPUT:TECH1 TECH10 TECH11 TECH14;
SAVEDATA:
FILE IS lca2turkey.dat;
SAVE IS CPROB;
FORMAT IS FREE;
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 336-347

Test Equating with the Rasch Model to Compare Pre-test and Post-test Measurements

Zeynep UZUN * Tuncay ÖĞRETMEN **

Abstract
The purpose of this study is to prove the equitability of pre and post-tests with the Rasch Model and to provide the observability of individual and interindividual ability changes by evaluating the equated tests with stack analysis within the scope of the Rasch Measurement Theory. The pre-test and post-test data that are applied in this study were derived from the project named A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model No. 115K531, which started on 15/11/2015 and was supported by the TUBİTAK SOBAG 3501 program. The tests were analyzed with the Rasch model, and the fit of the data to the Rasch model was evaluated, and then the Rasch Model and the Separate estimation-Common person method were applied for equating process. Lastly, individual and interindividual ability changes were observed by applying the stack analysis method with the Rasch model. As a result of the analysis of pre and post-tests with the Rasch model, it was concluded that they meet the requirements of the model. As a consequence of the equating process, the equitability of pre-test and post-test was proved, and it was observed that the individual and interindividual ability change could be evaluated by analyzing the pre-test and post-test data with the stack analysis method.

Key Words: Rasch model, test equating, stack analysis.

INTRODUCTION

When there is a need to compare tests, the first thing to be examined is whether the tests in question are comparable or not. For this purpose, the tests are equated with the equating methods, and, in the result of success, the comparability of the tests is proven.

Equating is defined as adjusting one test form's unit system to another test form's unit system (Angoff, 1971) and is a statistical process (Kolen & Brennan, 2004). It is applied with two methods: horizontal and vertical equating.

The horizontal equating is used at comparable difficulty levels and in need of equating the test forms in which the ability distributions of the candidates who take the exam are similar, while the vertical equalization is used at different difficulty levels and in need of equating the test forms in which the ability distributions of the candidates who take the exam are different (Hambleton & Swaminathan, 1985).

For instance, while the horizontal equating is used when the application of different forms of the test is required, the vertical equating can be used for the purposes such as; evaluating a student who performs well above her/his class with a test a few levels ahead, tracking learning development of a student with exams, evaluating multiple groups at different levels with a single scale (Hambelton & Swaminathan, 1985), working with standardized tests (Crocker & Algina, 1986), analyzing the effect of intervention as an individual by proving the comparability of scores, considering the possibility that in pre-test/post-test applications, which is also examined in this study, items may not function in the same way for all those who took the test (Anselmi, Vidotto, Bettinardi, & Bertolotti, 2015).

* Graduate Student, Ege University, Faculty of Education, Izmir-Turkey, zuzun2204@gmail.com, ORCID ID: 0000-0003-4681-0044
** Ph.D, Ege University, Faculty of Education, Izmir-Turkey, tuncay.ogretmen@ege.edu.tr, ORCID ID: 0000-0001-7783-1409

To cite this article:
Uzun, Z., & Öğretmen, T. (2021). Test equating with the Rasch model to compare pre-test and post-test measurements. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 336-347. doi: 10.21031/epod.957614

Received: 25.06.2021
Accepted: 29.11.2021

Equating can be applied based on two approaches: Item Response Theory or Classical Methods (Hambelton & Swaminathan, 1985). Classical Methods are divided into two as Equipercentile equating and Linear equating (Angof, 1971).

In the Equipercentile equating method, the scores in the two tests to be compared are accepted as equivalent if the frequency distributions are the same for a particular sample. The method ensures that the converted score distributions are the same. However, using raw scores causes problems in meeting the requirements, which were defined by Hambleton and Swaminathan (1985), such as subject independence, unidimensionality, and symmetry. Therefore, the Equipercentile equating method is considered group-dependent (Hambleton & Swaminathan, 1985).

In Linear equating, scores corresponding to the same standard score are considered to be equal (Angoff, 1971). Just like Equipercentile equating, it is group-dependent and does not meet the equating requirements (Hambleton & Swaminathan, 1985). In addition to that, in equating with classical methods in the comparison of pre-test and post-test scores, the statistical significance and magnitude of the difference between the average scores obtained from both tests can be mentioned. Examining the individual development of the students or the individual development differences among the students is out of the question.

Item Response Theory (IRT), on the other hand, is advantageous compared to classical methods since item and ability estimations can be made independently from the sample. If the item response model fits the data, the requirements in the classical method will be met due to its equality, symmetry, and invariance features (Kolen, 1981).

When the pre/post-test applications are compared by equating the test forms with the IRT, it is possible to examine not only the change in the average scores of the sample but also the individual development of the students. Two things can be achieved by measuring change at the individual level; first, characteristics that can separate students based on whether or not they have shown any development, which can be used in future applications, and secondly, the degree of change in ability seen in cases where the effect desired to be evaluated differs due to individual differences of students (Anselmi et al., 2015).

In the Item Response Theory, true score and observed score equating methods are recommended for equating (Kolen & Brennan, 2004). In the true score equating method, the tests are equated at the θ ability levels. Therefore, for equating Concurrent Estimation (Lord, 1980) and Separate Estimation methods are used.

In the observed score equating method, the score distributions of the tests are estimated with the selected IRT model, and the scores are equated with the Equipercentile equating method (Kolen & Brennan, 2004). When the Item Response Theory approach is preferred to equate the tests, it is necessary to decide which IRT model will be used for data analysis before choosing the equating method.

Purpose of the Study

The purpose of this study is to equate the pre and post-tests that are prepared and applied within the scope of the project named A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model No. 115K531 (Başokçu et al., 2018), which started on 15/11/2015 and was supported by the TUBİTAK SOBAG 3501 program using the Rasch model based on the Item Response Theory and the Separate Estimation-Common Person Equating. Tests evaluate with stack analysis to ensure the observability of individual and interindividual ability changes.

Subproblems

For pre-test analysis: Do the pre-test data fit the Rasch model? Is the pre-test unidimensional? Does the pre-test have sufficient distinctiveness?

For post-test analysis: Do the post-test data fit the Rasch model? Is the post-test unidimensional? Does the post-test have sufficient distinctiveness?

For equating procedure: Can pre-test and post-test scores be compared? Can pre-test and post-test scores be converted into each other? Can the individual and interindividual change of the effect be analysed by evaluating the pre-test and post-test data on the same scale?

METHOD

In this study, all data analyses were conducted with the Dichotomous Rasch model. The Separate estimation-Common person method was applied to evaluate whether the pre-test and post-test measures were comparable. Pre-test and post-test were compared with stack analysis.

Instrument and Sample

The first identical 29 items of pre-test/post-test exam, which consists of 30 multiple choice items, prepared within the scope of project A Model Proposal to Increase Turkey's Success in the field of Mathematics in International Large-Scale Exams: Effectiveness of the Cognitive Diagnosis based Monitoring Model, constitute the measuring instrument of the research while a total of 1225 six-graders in 42 classes of 5 different schools in Izmir province constitute the sample of the exam.

Procedure

In this study, analyses are carried out in three steps. In the first step, the data obtained from the students' pre-test and post-test applications are analyzed separately with the Dichotomous Rasch model, and the fit of the data with the model and the statistical characteristics of the tests are evaluated. In the second step, the equating process is applied between the pre-test and post-test with the common person Separate estimation-Common person method, the pre-test as the reference. And in the third step, the observability of individual and inter-individual ability changes are evaluated with the Dichotomous Rasch model and with the pre-test and post-test data, which are proven to be equitable with each other by stack analysis method.

Rasch Analysis

Rasch analysis is a single parameter IRT model that estimates test items' parameters and the characteristics that are intended to be measured according to the possible answers for the items. The ability and parameter estimations are independent of the sample to which the test is applied. In Rasch analysis, knowledge is a function of the difference between person ability and item difficulty. As with the Guttman scale, it is assumed that the person will answer all items up to her/his ability level correctly. In the Rasch model, individuals and items can be positioned on the same scale, and using the Equation 1, which was used in a one-parameter logistic model, the probability of a person with θ ability level to correctly answer item i in a b_i difficulty is calculated in the model.

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}},\ i = 1,2,\dots,n \qquad (1)$$

The Rasch model is applied by choosing the appropriate one among the three based on the number of item categories and weighting: Dichotomous, Andrich Rating Scale Model (RSM) or Master Partial Credit Models. In this study, analyses are carried out with the Winsteps 3.92.1 program using the Dichotomous Rasch model. The Rasch model requires unidimensionality, local independence, monotonic rising, and non-intersecting item response functions. In order to obtain test and item statistics and to evaluate the validity and reliability of the test, Rasch-based item-response threshold ordering, fit of the data to the model, item difficulty and person ability, unidimensionality and local independence, differential item functioning, scatter, and reliability analyses should be performed.

Item-category average measures

In the Rasch model, the values of the ability means corresponding to the categories are examined in order to evaluate the valid discriminability of the categories and to reveal if the item is understood correctly by the test takers. For a Dichotomous model, if the ability means of category 0 of an item is lower than the ability of category 1, it is established that the item was correctly understood by the test takers. The difference between the means indicates the power of discrimination.

Model fit tests

By evaluating the concordance statistics, it is determined to what extent the items, the individuals, and the test fit the Rasch model. In the analyses performed with the WINSTEPS program, item fit and person fit are evaluated with INFIT and OUTFIT MNSQ (mean square) sizes and unit standard deviation values. If the INFIT and OUTFIT MNSQ (mean square) sizes are between 0.5 and 1.5, it indicates that the scale is unidimensional and the sample size is sufficient. (Linacre, 2016)

In the present study, Among the 1225 students who took the pre-test and post-test, responses of 10 were excluded from the analysis due to missing data, as with the responses of a total of 4 more students, 2 in the pre-test and 2 in the post-test, were excluded from the analysis as well because their MNSQ values were higher than 4.0.

Item difficulty and person ability

In the Rasch model, item difficulty and person ability are expressed in logit. The difficulty of items refers to the corresponding level of ability. An item has a 50% probability of being answered correctly at the corresponding level of ability. A higher logit represents a more difficult item and a person with greater abilities.

Unidimensionality and local independence

The Rasch model requires meeting the unidimensionality assumption (Chang, Wang, Tang, Cheng, & Lin, 2014). Meeting the unidimensionality assumption is also an indication of local independence. Item parameters may be estimated biased in the state of unachieved unidimensionality under the unidimensionality assumption.

Differential item functioning

In order to evaluate whether the items show bias or not in the Rasch analysis, the size of the DIF contrast and the statistical significance of the Mantel Hanzel Chi-square value are examined. DIF contrast should be between -0.50 and 0.50 logit values, and Mantel Hanzel Chi-square value should be statistically insignificant ($p \geq .05$). A negative DIF contrast value indicates that the item is easy for the subject, while a positive DIF contrast value indicates that it is difficult for the subject (Linacre, 2016). Item bias can be seen as uniform item bias, where bias is seen at the same rate at all levels of ability, or as non-uniform item bias, where it occurs at specific or varying ability ranges. In this study, uniform item bias is evaluated based on the gender variable.

Separation and reliability

In the Rasch analysis, reliability is evaluated through personal reliability, person separation index, item reliability and item separation index. In the case of the measurement error getting smaller, the reliability values become insensitive and cannot exceed the upper limit of 1. At this point, the separation indices provide this congestion to be stated (Wright, 1996a).

Person separation is used in the classification of individuals, and its low value (< 2 and person reliability < .8) shows that the measuring tool is not sensitive enough to distinguish individuals at lower and upper-performance levels (Linacre, 2016).

Item separation, on the other hand, enables the evaluation of the concordance of item hierarchy to the expectations. Low item separation (< 3 = high, medium, low item difficulty, and item reliability < .9) indicates that the sample is not large enough to evaluate the rate of concordance between item hierarchy and expectations (Linacre, 2016).

Equating Treatment

When evaluating a certain effect with pre/post-test, to observe the change in individuals or the functioning of the items, pre and post-test should be equated. For this purpose, in the study, it was evaluated whether the data were suitable for stack analysis with separate estimation common person equating. Ability measurements obtained by analyzing the pre-test and post-test separately using the Dichotomous Rasch Model in Winsteps 3.92.1 program were used in the equating process. The process consists of three steps:

1. Using the ability measures obtained by the Dichotomous Rasch model, a trend line is obtained by placing the pre-test ability measures of a small portion of the sample on the x-axis and the post-test ability measures on the y-axis. If the line angle is 45 degrees to the x-axis, it is considered that the pre-test and post-test measures are convertible to each other, and the data are considered to be suitable for both stack analysis that allows the examination of change on an individual basis and rack analysis that allows examination of change on an item basis. If the trend line cannot provide the 45-degree angle,
2. Empirical intercept and slope values of the trend line are used to capture the slope. These values convert y-axis measures to x-axis measures or the reverse. If the intervention method whose effect is to be examined is to be evaluated, the post-test data is shifted by using the coordinate where the trend line cuts the x-axis and the slope value (pre-test parameters), and the trend line is obtained again. If the aim is to make a decision about the result of the desired intervention method, the pre-test should be shifted with the post-test parameters and the trend line is obtained again. If the new trend line cannot provide the 45-degree angle, it is assumed that equating is not possible between the two tests. If it provides the 45-degree angle, it is considered that the equating procedure is successful for the part taken from the sample, and to examine whether the equating will be valid for the whole sample,
3. Equating analysis is applied to entire sample data with the same coordinate and slope value. With the trend line obtained by using the whole sample, making an angle of 45 degrees with the x-axis, it is determined that the equating process is successful and the two test data are suitable for stack analysis.

In the present study, the ability measures of the first 68 students in the data set were used to evaluate the first step of the equating process. In the second step, the post-test data were shifted using the pre-test parameters.

Stack Analysis

To evaluate a specific effect applied and to evaluate the effect on an individual and group basis in the selected sample, where pre-test and post-test are applied, stack analysis is suggested (Wright, 1996b, 2003). Stack analysis is the analysis of the sample by combining the pre-test and post-test data. In this combination, the post-test data are added under the pre-test data as if different people took this exam. More specifically, stack analysis is Rasch analysis by arranging data. By adding the post-test data below the pre-test data, the data is stacked. Stack analysis is performed by applying Rasch analysis using stacked data. While the number of items does not change in the stacked analysis, the person sample doubles, and the difficulty of the items is kept constant between two-time points. To perform

stack analysis, equating pre and post-tests must be successful. In this study, stack analysis was applied with Winsteps 3.92.1 program using the Dichotomous Rasch model.

RESULTS

Rasch Analysis (First Step Results)

Item-category mean order

To examine if the pre-test and post-test item response categories were understood correctly by the students, the averages of the students who chose the 0 and 1 categories of each item were compared, and it was seen that the average ability values of the students who chose category 1 for each item were higher than the students who chose the category 0 of the item. The fact that the average ability values of the students who chose category 1 of items for the pre-test and post-test were higher than the students who chose category 0 of the item shows that the categories were correctly distinguishable, and the items were correctly understood. In other words, it was proven that the students can choose the categories that fit the purpose of the test.

Model fit tests

It was observed that the data of 1211 students included in the analysis fit the Rasch model (with an MNSQ value lower than 4.0). Students' mean infit and outfit values and standard deviations were revealed as follows: for the pre-test; mean infit 1.00 and SD 0.16, mean outfit 0.97 and SD 0.36, and for the post-test; mean infit 1.00 and SD 0.14, mean outfit 1.01 and SD 0.33. Since the infit and outfit values found were close to 1, it was stated that the sample fit the Rasch model.

When the infit and outfit values of the items were evaluated to evaluate the model fit, it was seen that the values in the pre-test and post-test were between 1.50 and 0.50, and the items were found to be fit with the model. The mean infit, outfit, and standard deviation values of the tests are revealed as follows: for the pre-test; mean infit 1.00 and SD 0.08, mean outfit 0.97 and SD 0.19, and for the post-test; mean infit 0.99 and SD 0.12, mean outfit 1.01 and SD 0.22. Since the infit and outfit values found were close to 1, it is stated that the test is compatible with the Rasch model.

Item difficulty and person ability

For the pre-test, the ability measurements ranged from 1.90 to -3.80 logit, and the average ability measurements were -1.38 (SD: 0.81) logit. The ability measurements for the post-test ranged from 5.01 to -3.67 logit, and the average ability measurement was -0.95 (SD: 1.11) logit. The item difficulty average measure for the pre-test was found to be 0.0, and the item difficulty average measure for the post-test was also found to be 0.0. The difficulty levels of the items as a result of the pre-test and post-test separate analyses and the difficulty values of the items found as a result of the equating process are given in Table 1 in The Stack Analysis section. In Figures 1 and 2, item difficulty and person ability distributions for pre-test and post-test are presented.

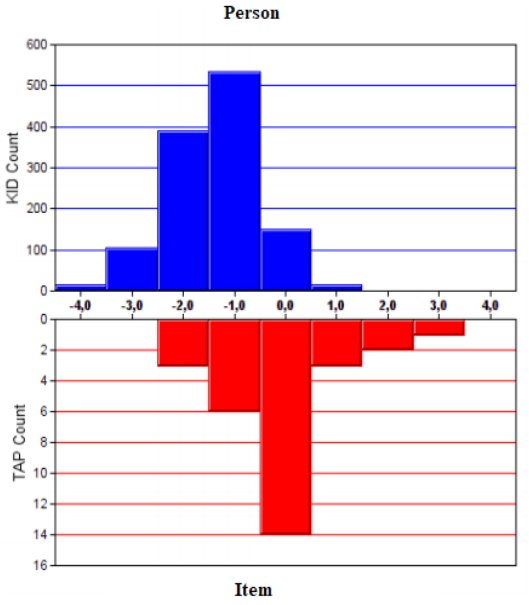

Figure 1. Pre-test İtem Difficulty and Person Ability Distribution

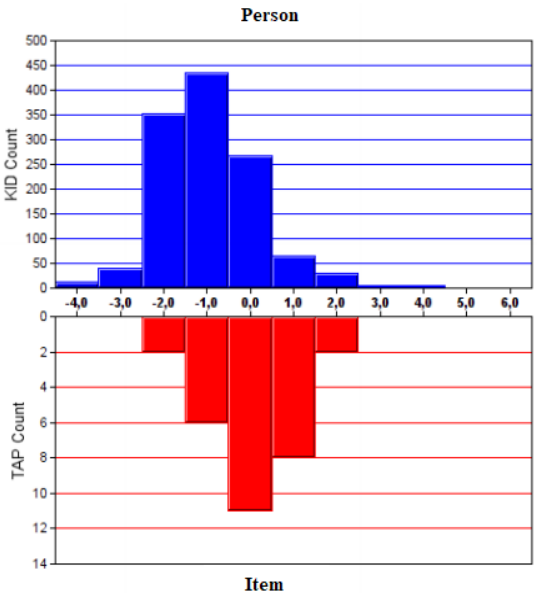

Figure 2. Post-test İtem Difficulty and Person Ability Distiribution

In Figure 1, the item range is between 3.03 and -1.87, while the skill range is between 1.90 and -3.80. In the pre-test, there is no item suitable for the ability level of the students evaluated on the logit 2 and 3 ability measure. As a result, it was determined that the pre-test could not differentiate the sample sufficiently.

Unidimensionality and local independence

For pre-test in the evaluation of multidimensionality, it was revealed that the percentage of observed explained variance (21.9%) and the percentage of unexplained observed variance (78.1%) were equal to the expected percentages; whereas the unexplained variance in the 1st contrast values was calculated as 1.78, revealing that the corresponding unexplained observed variance percentage was lower than the expected unexplained variance percentage. It can be said that there is no such doubt for the pre-test since it was considered as a multidimensionality possibility when the unexplained variance in the 1st contrast value is higher than 2.

As the results of the main component analysis, the Winsteps program divides the items into 3 clusters based on their 1st contrast loads and checks whether they measure the same thing by comparing them. The disattenuated correlation value difference between the item clusters being lower than 0.7 and the Pearson correlation value being lower than 0.3 indicates a second dimension. In the case of a second dimension, person measurements become biased.

It was determined that, in this study, for pre-test, the disattenuated correlation between item clusters 1 and 3 was lower than 0.7, and the Pearson correlation values between 1-3rd and 2-3rd clusters were lower than 0.3. The eigenvalue of significant contrasts between the items is greater than 2 (Linacre, 2016). It was determined that the multidimensionality effect did not create a significant difference since the unexplained variance contrast value for the pre-test was lower than 2.

For the post-test, the following was noted: observed variance percentage (26.3%) was higher than the expected value, the unexplained observed variance percentage (73.7%) was very close to the expected value (74%), and the unexplained variance in the 1st contrast value was calculated as 1.81, whose corresponding unexplained observed variance percentage was lower than expected unexplained variance percentage. It can be said that there is no such doubt for the post-test since it was considered as a multidimensionality possibility when the unexplained variance in the 1st contrast value is higher than 2. And since it was determined that the disattenuated correlations between item clusters were

greater than 0.7 and the Pearson correlation values were greater than 0.3, no doubt would suggest multidimensionality.

Differential item functioning

As a result of the DIF analysis performed to understand whether the items show gender bias, only the DIF contrast of the 24th item for the pre-test was found to be higher than 0.50, and the Mantel Hanzel Chi-square values ($p \geq .05$) were not statistically significant revealing that there was no gender bias. Since there was no DIF contrast higher than 0.50 for the post-test, likewise, it also means that the post-test did not show gender bias as well.

Separation and reliability

Based on the results, the Cronbach Alpha reliability coefficient of the pre-test analysis was .64, the reliability (Model) value of the individuals was .62, and the separation coefficient was 1.27. If the reliability value is below 0.80, it indicates that people are clustered in groups 1 or 2 (Linacre, 2016). When the individual reliability value is evaluated with the separation coefficient, the individual reliability and discriminability of the pre-test were found insufficient. Cronbach Alpha value was also found to be at an insufficient level. The item reliability coefficient and discrimination index of the pre-test were determined as .99 and 12.34, respectively, and the reliability and discrimination of the items were stated as quite good.

Based on the results, the Cronbach Alpha reliability coefficient of the post-test analysis was 0.83, the reliability (Model) value of the individuals was 0.80, and the separation coefficient was 2.02. The person reliability value between 0.80 and 0.90 indicates that the sample can be divided into 2 or 3 groups (Linacre, 2016). When the person reliability value was evaluated with the separation coefficient, the person reliability of the post-test was found sufficient. Moreover, Cronbach Alpha value was found to be at an insufficient level. The item reliability coefficient and discrimination index of the post-test were determined as .99 and 12.06, respectively, and the item reliability was stated as quite good. It was determined that the pre and post-tests meet the requirements and can be equated with the Rasch model approach.

Separate Estimation Common Person Analysis (Second Step Results)

The equating procedure with separate estimation common person analysis was applied to compare the pre-test and post-test data on the same metric. Considering that the intervention method that is being examined within the scope of the research will be developed, the equating method was applied by using pre-test parameters. Based on the Dichotomous Rasch model, the measurements of the first 68 people, starting from the highest ability level, were drawn with the pre-test measurement values on the x-axis and the post-test person measurement values on the y-axis. Furthermore, it was observed that the measurements were not parallel. In this case, the rack analysis was found inappropriate to apply. A correction was performed in the post-test using the coordinate -1.53, which is the point where the line obtained intersects the x-axis, and 0.73, which is the slope value of the line, to ensure the equating of the measurements. It was seen that the new line slope obtained was .997, and the equating process was found to be successful for 68 people. To examine whether the equating obtained as a result of the correction process will be valid for the whole sample, analysis was once more applied, this time considering the whole sample. The line slope was calculated as .996 and the equating between the two tests was still valid. The correlation value between tests was 0.52, and the common variance was 27%. The correlation value, free of measurement error, was calculated as 0.74. With this determination, it was proved that pre-test and post-test can be evaluated on the same metric with stack analysis and test scores can be converted to each other.

Equation 2 and Equation 3 can be used for conversion:

$$\text{Pre-test Score (x-coordinate)} * \text{slope} + \text{y-coordinate} = \text{Estimated Post-test Score} \quad (2)$$

Post-test Score (y-coordinate) / slope + x-coordinate = Estimated Pre-test Score (3)

In the scope of this study pre and post-test scores can be converted into one another using the Equation 4 and Equation 5:

Pre-test Score (x-coordinate) * 1.37 + 2.1 = Estimated Post-test Score (4)

Post-test Score (y coordinate) / 1.37 - 1.53 = Estimated Pre-test Score (5)

Stack Analysis (Third Step Results)

Stack analysis is the analysis of the sample by combining the data of the pre-test and post-test. In this combination, the post-test data are added under the pre-test data as if different people took this exam. While the number of items does not change, the person sample doubles and the difficulty of the items is kept constant between two-time points. For the stack analysis applied, it is evaluated whether the item categories are understood according to the purpose of the test. When the average of the response categories of the items was examined, it was determined that the average ability value of the students who preferred the category 1 of the item, for all items, was higher than the students who chose the category 0. This situation means students can choose the categories according to the purpose of the test. In Table 1, stack Analysis, pre-test, post-test item difficulty measurements, and mathematical general and domain-specific ability areas are given.

Table 1. Pre-test, Post-test, and Stack Analysis Item Measurements

Item	Content Domain	Mathematical Abilities				Pre-test	Stack Analysis	Post-test	Level Change of Difficulty
		Mathematization	Using symbolic and technical language	Reasoning and developing a strategy	Communication and association				
1	Number	X				-1.87	-1.69	-1.55	Harder
2	Number		X			-1.03	-1	-1.02	Same
3	Number	X		X		-1.06	-0.96	-0.9	Harder
4	Number		X		X	-0.66	-0.45	-0.29	Harder
5	Number	X	X			0.09	0.15	0.16	Harder
6	Number	X	X			0.4	0.53	0.62	Harder
7	Number	X		X		-1.23	-1.03	-0.86	Harder
8	Number		X			-1.79	-1.69	-1.64	Harder
9	Number	X	X			-0.45	-0.4	-0.39	Harder
10	Number	X	X		X	0.99	1.08	1.12	Harder
11	Number	X				1.78	0.35	-0.39	Easier
12	Number			X	X	0.14	0.08	-0.01	Easier
13	Number	X				0.39	0.24	0.09	Easier
14	Number	X	X			0.07	-0.25	-0.57	Easier
15	Geometry	X		X	X	0.1	0.39	0.63	Harder
16	Geometry	X		X	X	0.9	0.75	0.6	Easier
17	Number	X			X	-0.49	-0.11	0.24	Harder
18	Number	X				-0.52	-0.37	-0.27	Harder
19	Number	X		X		0.46	0.56	0.61	Harder
20	Number	X		X	X	0.98	0.99	0.97	Same
21	Number			X	X	-0.18	0.05	0.23	Harder
22	Number			X	X	0.15	-0.13	-0.41	Easier
23	Number			X		-1.61	-1.39	-1.2	Harder
24	Number	X	X	X	X	3.03	2.65	2.45	Easier
25	Number		X		X	0.46	0.6	0.69	Harder
26	Number	X	X			-1.04	-0.96	-0.94	Harder
27	Number	X		X		1.87	1.67	1.52	Easier
28	Number	X		X		-0.32	-0.21	-0.14	Harder
29	Number	X		X		0.41	0.56	0.65	Harder

In Table 1, item difficulties are expressed as logit, and changes in item difficulty in the post-test compared to the pre-test are indicated in the difficulty level change column in order to make it easier

to notice. When the items were examined, it was found that the item difficulty values of the items 1, 3, 4, 5, 6, 7, 8, 9, 10, 15, 17, 18, 19, 21, 23, 25, 26, 28, and 29 increased in the post-test compared to the pretest. The pretest difficulty for the item 2 was -1.03 logit, and the post-test difficulty was -1.02 logit, and for the item 20 the pre-test difficulty was 0.98 logit, and in the post-test 0.97 logit and it was determined that the item difficulty values were very close. On the other hand, it is seen that the difficulty values of the other items decreased in the post-test compared to the pretest; in other words, they were easier for the students. In order to inform about the content of the items, the general and domain-specific mathematical abilities of the items, which are prerequisites in mathematical literacy (Başokçu et al., 2018), are also included in Table 1.

Item difficulty values can't change between pre-test and post-test in stack analysis. Items get a fixed value for two tests. The cross graph of the pre-test and post-test ability measurements of the stack analysis, which shows the individual ability change when the item difficulties are kept constant at two-time points, is given in Figure 3.

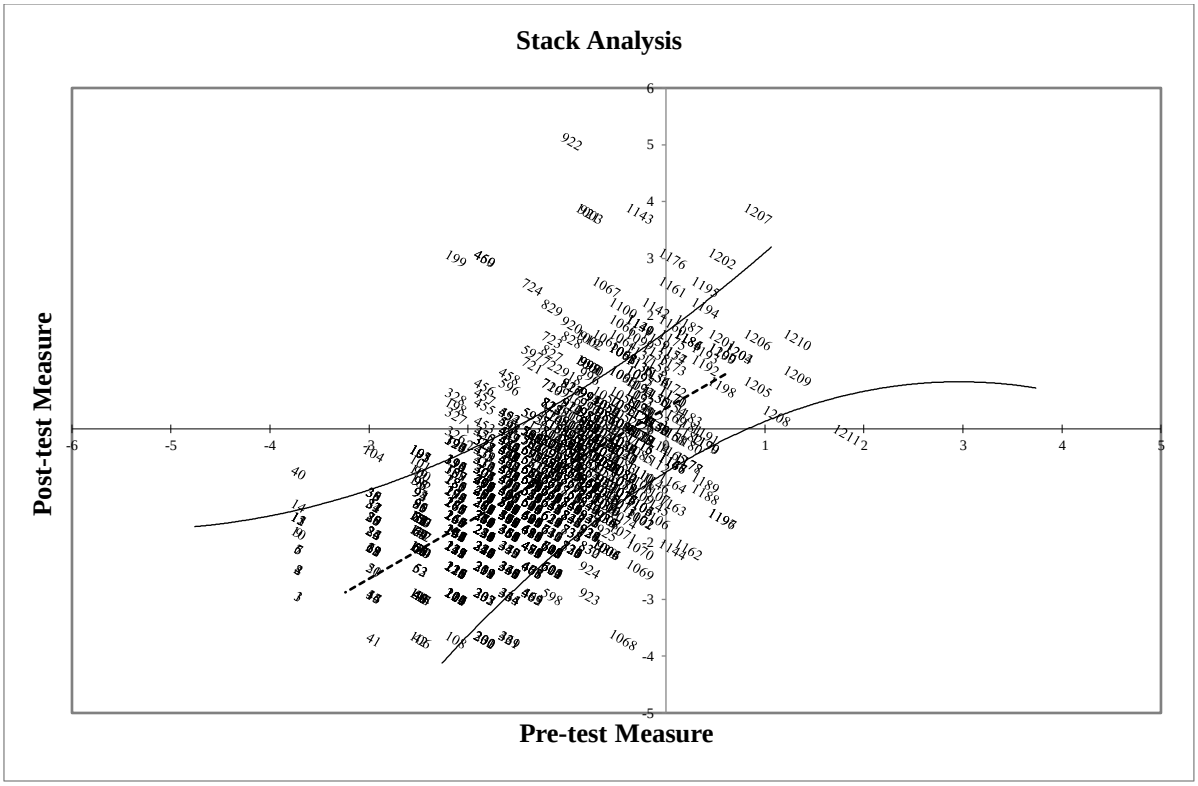

Figure 3. Cross Chart of Pre-test and Post-test Ability Measurements of Stack Analysis

In Figure 3, the pre-test measures of the students are shown in the x-axis and the post-test measures in the y-axis are shown as logit. There is a statistically significant difference ($t_{(1210)} = 12.79$, $p < .05$) in favor of the post-test between the pre-test and post-test measures results of the students. The correlation between the pre- and post-test measures of the stack analysis was 0.52. The correlation was found to be moderate. This shows that the effect applied between the pre-test and the post-test leads to different levels of change among students. Students above the identity line performed better in the post-test than the pretest. For a student who falls below the identity line, the situation is the opposite. In the graph, it is seen that the success of students numbered 922, 199, 459, and 460 increased more than other students. The ability measurement of student number 922 was calculated as -0.95 logit in the pre-test and 5.06 logit in the post-test, and an ability increase of 6.01 logit was observed. This value is 5.12 for the student numbered 199, 4.83 for the student numbered 459 and 460. Students numbered 1068, 1162, and 1069, which are below the identity line and at the farthest point, were negatively affected by the

effect between the pre-test and the post-test, and a decrease in the ability of 3.23, 2.35 and 2.22 logits was observed in the students, respectively.

DISCUSSION and CONCLUSION

When the results of item bias, multidimensionality, and discrimination analysis of the post-test are compared with the pre-test results, it is seen that the pre-test results were much weaker than the post-test results. The reason for this situation may be that students encounter types of questions that they did not encounter before in the pre-test application, while this effect disappeared in the post-test since they grasped the structure of the item better through the follow-up tests. One of the project findings from which the data was collected supports this view. This finding is that the problem situations that students encounter in the skill area they want to gain affect their success. Exposing students to problem situations similar to those in the tests aimed at increasing the level of success will increase success (Başokçu et al., 2018).

In the equating procedure, it has been proven that the pre-test and post-test ability measures are convertible to each other, as the slope of the trend line obtained with all sample ability measures in the pre-test and post-test to the x-axis is .996. Within the scope of this study, there was no need to transform ability measures with conversion formulas. The convertibility has only been demonstrated, as the ability for stack analysis needs to be convertible between pre- and post-testing of the ability measures. It has been observed that the ability measures obtained from the pre-test and post-test are comparable and can be evaluated on the same scale as the stack analysis.

In accordance with the previous study (Başokçu et al., 2018), it was observed that there was a significant difference in favor of the post-test between the pre-test and post-test ability measurements obtained within the scope of the stack analysis. As a result of the comparison of the pre-test and post-test ability measures obtained with the stack analysis with the help of graphics, students who were differently affected by the effect applied between the pre-test and the post-test could be determined. The change in students' ability levels can be compared. Thus, it has been seen that the level of individual and inter-personal ability change can be evaluated.

In the pre- and post-test evaluation, there are studies in which common item equating is used (e.g. Fujita & Mayekawa, 2011) or only the stack analysis method is used without using the equating procedure (e.g. Cunningham & Bradley, 2010; Herrmann-Abell, Flanagan, & Roseman 2012; Ling, Pang & Ompok, 2018). Common person equating was preferred to prove the equivalence of equivalent forms structure (e.g. Cavanagh, 2012; Popp & Jackson, 2009; Taylor & McPherson, 2007). It was stated by Masters (1985) that the same results can be achieved with both equatings in the Rasch Model, and the common person equating tests unidimensionality more clearly. For this reason, common person equating and stack analysis are used in this study to show that the intervention effect can be evaluated on an individual basis. Compared to previous studies, a stricter 1st contrast value was taken as the criterion in this study compared to Ling et al. (2018) study. Compared to the studies of Cunningham and Bradley (2010) and Anselmi et al. (2015), it is presented with a better percentage of person who fit the model. The results are generally in accordance with previous studies, and no feature has been identified that can make the procedure specific or hinder the implementation of equating or stack analysis.

In this study, the factors that affected the students who benefited more from the effect applied between the pre-test and the post-test or could not benefit from it were not researched. The factors that increase or decrease the success of the student can be determined by interviewing the students who are affected differently individually or by applying tests on possible factors to these students. Thus, the method of intervention can be developed, individualized, or differentiated for the groups to be determined. In the field of education, each student's unique and individual talent is an investment for the future of society. From this point of view, it is considered that the equating steps with the Rasch model are a suitable choice for studies that evaluate the intervention methods (the effect of the use of materials, the effect of the teaching model, etc.) whose effects are desired to be investigated with the pre-post test application.

REFERENCES

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Anselmi, P., Vidotto, G., Bettinardi, O., & Bertolotti, G. (2015). Measurement of change in health status with Rasch models. *Health and Quality of Life Outcomes, 13*(1), 1-7. doi: 10.1186/s12955-014-0197-x

Başokçu, O. T., Bardakçı, V., Çakıroğlu, E., Öğretmen, T., Yurdakul, B., & Akyüz, G. (2018). *Uluslararası geniş ölçekli sınavlarda Türkiye'nin matematik başarısını arttırabilmek için bir model önerisi: Bilişsel tanıya dayalı izleme modelinin etkililiği* (Proje No. SOBAG 3501). Retrieved from https://open.metu.edu.tr/bitstream/handle/11511/50310/TVRnMU56SXk.pdf

Cavanagh, R. F. (2012, December). *Engagement in classroom learning: Ascertaining the proportion of students who have a balance between what they can do and what they are expected to do.* Paper presented at the 2012 Annual International Conference of the Australian Association for Research in Education, Sydney, Australia.

Chang, K. C., Wang, J. D., Tang, H. P., Cheng, C. M., & Lin, C. Y. (2014). Psychometric evaluation, using Rasch analysis, of the WHOQOL-BREF in heroin-dependent people undergoing methadone maintenance treatment: Further item validation. *Health and Quality of Life Outcomes*, *12*(1), 1-9. Retrieved from https://link.springer.com/article/10.1186/s12955-014-0148-6

Crocker, L., & Algina J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cunningham, J. D., & Bradley, K. D. (2010, May). *Applying the Rasch model to measure change in student performance over time*. In American Educational Research Association Annual Meeting, Denver, CO.

Fujita, T., & Mayekawa, S. I. (2011). A comparison between common item equating with pre-and post-reading and listening tests. In C Ho, M.-F. G. Lin (Eds.), *E-Learn: World conference on e-learning in corporate, government, healthcare, and higher education* (pp. 626-631). Association for the Advancement of Computing in Education (AACE).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Academic Publishers Group.

Herrmann-Abell, C. F., Flanagan, J. C., & Roseman, J. E. (2012, March). Results *from a pilot study of a curriculum unit designed to help middle school students understand chemical reactions in living systems* (Online Submission). Paper presented at the NARST Annual International Conference, Indianapolis.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1-11. Retrieved from http://www.jstor.org/stable/1434813

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York: Springer.

Linacre, J. M. (2016). *Winsteps® (Version 3.92. 0)* [Computer Software]. Winsteps, Beaverton, OR, USA. Retrieved from www.winsteps.com

Ling, M. T., Pang, V., & Ompok, C. C. (2018). Measuring change in early mathematics ability of children who learn using games: Stacked analysis in Rasch measurement. In Q. Zhang (Ed.), *Pacific rim objective measurement symposium (proms) 2016 conference proceedings* (pp. 215-226). Singapore: Springer.

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

Masters, G. N. (1985). Common-person equating with the Rasch model. *Applied Psychological Measurement, 9*(1), 73-82. doi: 10.1177/014662168500900107

Popp, S. E. O., & Jackson, J. C. (2009, April*). Can assessment of student conceptions of force be enhanced through linguistic simplification? A rasch model common person equating of the FCI and the SFCI*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Taylor, W. J., & McPherson, K. M. (2007), Using Rasch analysis to compare the psychometric properties of the Short Form 36 physical function score and the Health Assessment Questionnaire disability index in patients with psoriatic arthritis and rheumatoid arthritis. *Arthritis Care & Research, 57*(5), 723-729. doi: 10.1002/art.22770

Wright, B. D. (1996a). Reliability and separation. *Rasch Measurement Transactions*, *9*(4). Retrieved from https://www.rasch.org/rmt/rmt94n.htm

Wright, B. D. (1996b). Time 1 to time 2 (pre-test post-test) comparisons and equating: Racking and stacking. *Rasch Measurement Transactions, 10*(1). Retrieved from https://www.rasch.org/rmt/rmt101f.htm

Wright, B. D. (2003). Rack and Stack: Time 1 vs. time 2 pre-post. *Rasch Measurement Transactions, 17*(1), 905-906. Retrieved from https://www.rasch.org/rmt/rmt171a.htm

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 348-357

Comparison of Kernel Equating and Kernel Local Equating in Item Response Theory Observed Score Equating

Merve YILDIRIM SEHERYELİ * Hasibe YAHSİ SARI ** Hülya KELECİOĞLU ***

Abstract
The present study aims to compare the Kernel equating and Kernel local equating methods in observed score equating. Functions and error estimates regarding the difference between raw and equated scores and the scores equated by Stocking-Lord and Haebara true-score equating methods in Kernel local equating and Kernel equating were examined in Item Response Theory Observed Score Equating. Therefore, 5, 10, and 15 external anchor items were used, and scores were obtained from two forms based on the 2PL model. R (version 3.5.3.) programming software was used for IRT assumptions, item parameters, calibration, and equating analyses. The results revealed that Stocking-Lord and Haebara true-score equating methods yielded similar results. Moreover, if the equating method is the same, estimation errors decreased when the number of anchor items increased. The mean scores obtained by Kernel equation 5 and 15 anchor items were lower than Kernel local equating, while means of Kernel equating of 10 anchor items were higher. As the number of items increased, estimation errors decreased, and Kernel local equating revealed the lowest errors in the medium score scale. Kernel equating can be used based on the related ability level if the individual's ability distribution is known.

Key Words: Test equating, Kernel equating, Kernel local equating, item response theory, local equating.

INTRODUCTION

Measurement tools are used for many purposes, such as measuring cognitive, affective, or psychomotor characteristics of individuals, getting to know individuals, placing them in any institution or school. The validity and reliability of the measurements of these measurement tools are a need for better measurement. To increase test reliability and therefore test validity, different test forms measuring the same feature are used especially in exams with wide participation and high risk, such as selection and placement exams, whose results greatly affect the future of individuals. These different test forms must have the same degree of difficulty for individuals to be evaluated fairly (Haladyna & Downing, 2004). However, this is not always possible in practice. In this case, the scores obtained from the test forms that do not have the same difficulty level should be brought to the same scale and the scores should be equated so that the forms can be used interchangeably. These statistical processes are possible with the help of test equating (Kolen & Brennan, 2004). Equating brings the scores on the test forms to the same scale, allowing them to be used interchangeably and the scores to be compared (Hambleton & Swaminathan, 1985). Thus, bias towards the measurement tools used in different test forms can be eliminated.

Equating has certain steps to be followed. The first step is to determine the data collection design. There are five data collection designs, which include equivalent groups design, single group design, counterbalanced design, non-equivalent groups with anchor test design, and non-equivalent groups with covariates design (González & Wiberg, 2017). The second step is to determine the equating method. These methods differ based on classical test theory (CTT) and item response theory (IRT).

* Ress. Assist., Hasan Kalyoncu University, Faculty of Education, Gaziantep-Turkey, yldrm.mrv.7806@gmail.com, ORCID ID: 0000-0002-1106-5358
** PhD. Student, Hacettepe University, Faculty of Education, Ankara-Turkey, hsbyahsi@gmail.com, ORCID ID: 0000-0002-0451-6034
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yıldırım Seheryeli, M., Yahsi Sarı, H., & Kelecioğlu, H. (2021). Comparison of Kernel equating and Kernel local equating in item response theory observed score equating. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 348-357. doi: 10.21031/epod.900843

Received: 22.03.2021
Accepted: 30.09.2021

Diao (2018) categorizes CTT test equating methods into four: identity equating, mean equating, linear equating and equipercentile equating. CTT-based equating methods differ among themselves based on true-score equating and observed score equating. Before starting the equating process in IRT, separate and concurrent calibrations are conducted to scale parameters. Item parameters of the two forms are estimated at the same time in concurrent calibration. In separate calibration, on the other hand, forms are scaled separately and calibrated using common items. Calibration methods include the moment method (mean-mean, mean-sigma) and characteristic curve transformation (Haebara and Stocking-Lord) (Kolen & Brennan, 2004). At the last stage of equating, standard errors are calculated, and properties of equating are checked. These properties are symmetry, same specifications, equity, observed score, and group invariance (González & Wiberg, 2017; Kolen & Brennan, 2004).

There are equating studies based on non-equivalent groups with covariates in IRT. Also, the effect of skewness of ability distributions, multidimensionality, violation of group invariance in different levels on equating errors have been examined in some studies (Gök & Kelecioğlu, 2014; Öztürk-Gübeş, 2019; Öztürk-Gübeş & Kelecioğlu, 2015; Tanberkan-Suna, 2018; Uysal, 2014). The general results of the studies showed that IRT true score equating method performed best in providing test fairness, while IRT observed score equating method performed best in decreasing measurement errors.

Equapercentile and linear equating methods are used in observed-score equating (von Davier, 2008). These methods include equapercentile equating methods, linear equating methods, IRT observed score equating, local equating, non-linear equating, and Kernel equating (von Davier, 2013).

Kernel equating is defined as an equapercentile equating method that transforms discrete score distribution into a continuous distribution (von Davier, Holland & Thayer, 2004). Kernel equating was first defined by Holland and Thayer (1981). The Kernel equating methods can be applied as post-stratification equapercentile, post-stratification linear, chained equapercentile and chained linear. The equating methods based on CTT uses linear estimates for the continuation of the score distributions, while Gauss Kernel method is used in Kernel equating (von Davier et al., 2004). There are five steps in observed score Kernel equating: pre-smoothing, estimating score distributions for the target population, computing the equating function, continuizing the discrete score distributions, and computing the standard error of equating (von Davier, 2013). Pre-smoothing helps the data become consistent. Kernel equating smoothes data transformation and provides a small standard error. Also, it is less affected by the change in the sample compared to other methods. Kernel equating is used with equivalent groups design, single group design, counterbalanced design, and non-equivalent groups with anchor test design (von Davier et al., 2004). There are various studies that used Kernel equating (Akın Arıkan, 2017; Andersson & Wiberg, 2014; Choi, 2009; Liou, Cheng, & Johnson, 1997; Norman Dvorak, 2009; Wiberg, van der Linden, & von Davier, 2014). Akın Arıkan (2017) found that the extreme scores yielded greater standard errors as the group ability distributions varied in Kernel equating. In IRT true score equating; on the other hand, middle and high scores had the greatest error. Kernel equating methods had lower standard errors in the medium score scale and had higher standard errors in extreme scores where score frequency was lower compared to the IRT true score equating in all conditions. Moreover, lower errors were obtained through the IRT true score equating method than the Kernel equating methods regarding the extreme scores.

Local equating also became popular along with Kernel equating in observed score equating (von Davier et al., 2004). It was first introduced by Lord (1980) in his definition of equating (as cited in van der Linden, 2000). All traditional equating methods use the same equating transformations for all populations of test participants. van der Linden (2000) revealed that equating should be done separately for each ability level. Local equating offers a common ground for different transformations for each ability level. In local equating, if both test forms are appropriate for the item response theory (IRT) and can be used with any equating design, the IRT observed score could be defined as the local Kernel equating (Wiberg et al., 2014).

Wiberg et al. (2014) proposed three different observed score Kernel local equating methods by combining local equating and Kernel equating. The methods for local Kernel equating on-equivalent groups with anchor test design are: IRT observed score equating, anchor test score Kernel equating

and local Kernel equating with ability estimated by anchor test. These new methods are compared to previous methods in terms of measures such as bias and relative error percentage. The item response theory observed score local Kernel equating method, which is used for all common equating methods, yielded bias, relative error, and Kernel standard equalization error, even when the measurement precision of the test was reduced. Kernel local equating methods generally showed low bias in the non-equivalent groups with anchor test design. In addition, the anchor was highly stable against variations in the accuracy and length of the test.

Many studies used Kernel equating (Akın Arıkan, 2017; Andersson & Wiberg, 2014; Choi, 2009; Liou et al., 1997; Norman Dvorak, 2009; Wang, Zhang ve You, 2020; Wiberg et al., 2014). These studies revealed that Kernel equating and similar traditional equating methods can be compared with equivalent groups design in equivalent and non-equivalent groups with anchor test design when estimating standard errors in equapercentile equating (Choi, 2009); and that R program was used for IRT observed score Kernel local equating (Andersson & Wiberg, 2014). Wiberg et al. (2014) used three different observed score Kernel local equating methods by combining local equating and Kernel equating in their study. Studies have shown that with Kernel local equating, equating functions can be obtained at each ability level, and thus estimation errors can be minimized (González & Wiberg, 2017; Wiberg et al., 2014). The present study compares the Kernel local equating with Kernel equating to examine the bias in the equating processes and the contribution of the methods to the test validity.

In kernel equating methods, in cases where the ability distributions between groups are different, extreme scores yield high standard errors (Akın Arıkan, 2017). Wiberg et al. (2014) concluded that in the common item nonequivalent groups, It is predicted that Kernel local equating methods will yield a lower standard error in cases where the ability distributions of individuals are known and the test fairness is ensured to make more accurate equating. In this study, the results of Kernel local equating are compared with Kernel equating under various conditions. Since there are few studies on this subject (Akın Arıkan, 2017; Wiberg et al., 2014) and there are no studies that examine Kernel equating and Kernel local equating together, the study aims to compare the results of Kernel equating and Kernel local equating.

Purpose of the Study

In the present study, θ values with values decreasing one by one between -6 and 0 (low), θ = 0 (middle) and θ values with values increasing one by one between 0 and +6 (high) ability levels of the scores obtained from two different forms based on 2PL model and different anchor item numbers in IRT observed score Kernel equating and IRT observed score Kernel local equating were included. Stocking-Lord and Haebara were used for data transformation and the equating results were compared. To this end, different anchor item numbers (10, 20, and 30) were used and after data were transformed with Stocking-Lord and Haebara methods, equating functions and errors were examined with observed score Kernel equating and observed score Kernel local equating.

METHOD

Research Design

In this study, data were artificially produced in order to examine the change of errors in cases where different anchor items were used in the equating methods and these items were not included in the total score. Therefore, this study is a simulation study.

Data Production

The items in the X and Y forms and the anchor materials were produced under the conditions in Table 1 according to the 2PL Model. The items in both data sets were produced using the “kequate” package

(Andersson, Bränberg & Wiberg, 2020) in the R Studio interface of the R (version 3.5.3) programming software with 20 items different and the number of anchor items 10, 20 and 30.

The forms consisting of X form and anchor items are named as P and Y forms and forms consisting of anchor items as Q forms. A parameters are uniformly distributed with values ranging from 0.50 - 2.0, the *b* parameters are *N*(0; 1) and the ability parameters are *N*(0.50; 1) for the P form, and *N*(0; 1) for the Q form.

According to Baker (trans. 2016), *a* parameters range between -2.80 and +2.80 in practice, while *b* parameters range between -3.00 and +3.00. In addition, he specified the cutoff point as 0.35 for the low level of the parameter *a*, 0.64 for the medium level, and 1.35 for the high level. In order to have medium and high level *a* parameters, they were taken between 0.50 and 2.00.

Wang, Lee, Brennan, and Kolen (2008) stated that the similarity distributions are important in terms of equating results and they considered the difference over 0.25 as *very wide*. Therefore, it was expected that the difference was taken as 0.50 to reveal the difference between Kernel equating and Kernel local equating errors more clearly.

Kolen and Brennan (2004) stated that the ratio of the number of anchor items to the total number of items in the test should be at least 20%. Therefore, while 30 items were different in all data sets (X and Y forms), the number of anchor items was determined as 5, 10 and 15. The number of iterations was determined as 100. A total of 600 (2 x 3 x 100) data sets were obtained. The average and ranges of the difficulty and discrimination parameters of the sample distributions obtained from the P and Q forms according to the number of anchor items are given in Table 1.

Table 1. The Mean and Ranges of the Difficulty and Discrimination Estimates Obtained From the P and Q Forms According to the Anchor Item Numbers

		P forms				Q forms			
Total Number of Items	Number of Anchor Items	Mean b	Range b	Mean a	Range a	Mean b	Range b	Mean a	Range a
35	5	-0.505	4.179	1.261	1.622	-0.013	4.454	1.253	1.614
40	10	-0.534	4.502	1.248	1.653	0.002	4.448	1.253	1.677
45	15	-0.537	4.625	1.255	1.656	-0.031	4.547	1.244	1.624

Table 1 shows that for each anchor item number, the means of *b* parameters related to the P forms are approximately 0.50 lower than the Q forms. The means of *a* parameter are approximately the same.

Data Analysis

In IRT Kernel equating (post-stratification equating) parameters were calibrated based on Stocking-Lord and Haebara methods. Then, an external anchor design was used in which the anchor items were not included in the total score. The distribution of equated scores, means of equating errors, and functions related to the difference between the raw score and the equated scores were compared regarding both calibration methods. Similar studies in kernel local equalization for cases where ability levels are low (decreasing one by one between -6 and 0) (L: Low), zero (M: Medium), and high (increasing one by one between 0 and +6) (H: High) was also repeated. “psych” (Revelle, 2021), “mirt” (Chalmers et al., 2021), “kequate” (Andersson et al., 2020), “ltm” (Rizopoulos, 2018) packages were used in the R Studio interface of the R (version 3.5.3) programming software for all analyzes.

RESULTS

This chapter presents the results of IRT Kernel observed score equating and Kernel observed score local equating. In this regard, score distributions, equating functions and distribution of equating errors were examined.

Equated Score Distributions and Functions of Score of Difference

In cases when the numbers of anchor items were 5, 10 and 15, Stocking-Lord and Haebara methods were examined. Ability levels for Kernel local equating are low (decreasing one by one from 0), medium (0), and high (increasing one by one from 0). Table 2 shows the equated score distributions obtained with the IRT observed score Kernel equating and Kernel local equating. These values were the result of calculating the means of the values obtained with 100 iterations.

Table 2. Equated Score Distributions Obtained With IRT Observed Score Kernel Equating and Kernel Local Equating

		Kernel Equating					Kernel Local Equating			
Calibration	Number of Anchor	Min.	Max.	Mean	S.D.	θ level	Min.	Max.	Mean	S.D.
Stocking-Lord	5	0.140	29.905	15.054	8.989	L	0.141	29.907	15.071	9.000
						M	0.167	29.906	15.039	9.022
						H	0.173	29.905	15.023	9.008
	10	-0.026	29.917	14.827	9.067	L	-0.022	29.782	14.817	9.039
						M	-0.048	29.787	14.792	9.060
						H	-0.045	29.916	14.802	9.086
	15	0.112	29.876	15.008	9.005	L	0.112	29.944	15.044	9.039
						M	0.069	29.948	15.009	9.070
						H	0.073	29.877	14.973	9.033
Haebara	5	0.141	29.912	15.067	8.993	L	0.142	29.919	15.084	9.005
						M	0.181	29.919	15.058	9.021
						H	0.186	29.912	15.042	9.007
	10	-0.032	29.921	14.824	9.073	L	-0.026	29.782	14.812	9.043
						M	-0.051	29.787	14.790	9.062
						H	-0.049	29.921	14.802	9.090
	15	0.117	29.871	15.002	8.997	L	0.118	29.938	15.040	9.033
						M	0.064	29.942	15.001	9.069
						H	0.068	29.871	14.963	9.030

Note. L: Low. M: Medium. H: High

Kernel equating results in Table 2 shows that when 10 anchor items were used in both calibrations, the equated scores were estimated with a low mean score. Again, in both estimation methods, the condition in which scores are estimated with a higher mean is the case where the number of anchor items is 5. The number of anchor items shows that the mean and standard deviations of the scores equated according to the methods do not differ much. Figure 1 shows the function graph regarding the differences between the equated scores and the raw scores taken from the test.

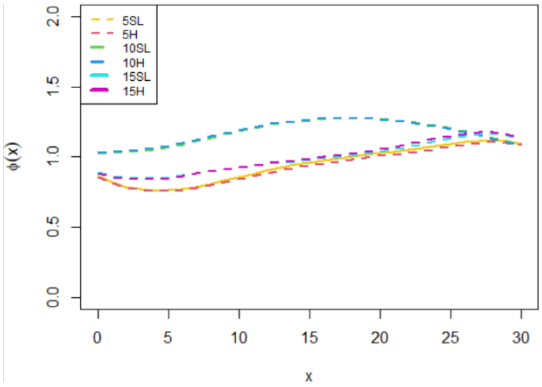

Figure 1. Function Graphs Regarding the Differences of Equated Scores and Raw Scores Obtained With IRT Kernel Equating

Figure 1 shows that the distribution of the difference scores is almost the same when the number of anchors is the same and the calibration method is different. When using 5 anchors, the differences decreased up to approximately 5 raw points, while the differences gradually increased after 5 raw points. When 10 anchor items were used, the difference scores increased as the raw scores increased and started to decrease after about 17 raw scores. When 15 anchor items were used, the difference scores increased as the raw scores increased, but the rate of change was relatively low. On the other hand, when 10 anchor items are used up to 26 raw points, the differentiation from the raw scores in both methods is higher compared to the other anchor items. When the raw score is greater than 26, it is seen that the differentiation from the raw score is more when 15 anchor items are used.

Table 2 demonstrates the Kernel local equating results and reveals that the mean scores of the equated scores at low, medium, and high ability levels are the highest in 5 anchor items and the lowest in 10 anchor items when Stocking-Lord (S-L) and Haebara (H) methods are used. In the case where 5 anchor items were used, the highest mean score was obtained in the low ability level with Haebara method, while the lowest mean score at the high ability level was obtained with the S-L method. When 10 anchor items were used, the highest mean score was obtained at the low ability level with S-L, and the lowest mean score at the medium ability level was obtained with the Haebara method. In the case where 15 anchor items were used, the highest mean score was obtained at the low ability level with S-L, and the lowest mean score at the high ability level was obtained with the Haebara method. In all conditions, the lowest mean score was obtained with the middle ability level when 10 anchors and the Haebara method was used, and the highest mean score was obtained with the low ability level when 5 anchors and the Haebara method were used.

When both methods are compared, in the case that 5 anchor items were used, the mean score obtained with Kernel equating was lower than the mean score obtained with Kernel local equating based on low ability level. The closest mean score was obtained when the Haebara method is used with the middle ability level. In the case where 10 anchor items are used, the mean scores obtained with Kernel equating are higher under all equating conditions. The closest mean score was obtained when the S-L method is used with the low ability level. In the case where 15 anchor items are used, the mean score obtained with Kernel equating is lower than the mean score obtained with Kernel local equalization with the low ability level. The closest mean score is the case in which the S-L method is used in the equating made according to the middle ability level. Figure 2 shows the function graph regarding the differences between the equated scores and the raw scores obtained from the test.

Figure 2 reveals that the distribution of the difference scores is almost the same when the number of anchors is the same and the calibration method is different. In the case of using 5 anchor items, the difference scores are higher in the equalizations made according to medium and high ability levels up to 14 raw points, while the difference scores are higher in the equations made according to the high ability level and Kernel equating over 17 raw points. In cases where 10 and 15 anchor items are used, up to 12 raw points, the difference scores are higher in the equalizations made according to medium and high ability levels, while the difference scores are higher in the equalizations made at medium and low ability levels over 23 raw points. In the equating made according to the middle ability level, the range of difference scores in each anchor item condition is the smallest.

Error Distributions

Equating errors were calculated for all conditions. These values were the result of calculating the means of the values obtained with 100 iterations. Table 3 shows the distribution of equating errors obtained with the observed score Kernel equating and Kernel local equating.

Table 3 shows that the error means of equated scores under all conditions are estimated higher in the Stocking-Lord method than in the Haebara method. The difference between these error means was found to be approximately .004 when 5 anchor items were used, .002 when 10 anchor items were used, and .003 when 15 anchor items were used. The distribution of the errors shows that as the number of anchor items increases, the errors are closer to each other and become more homogeneous. The

smallest of these errors occurred in the calibration made according to both methods when 15 anchor items are used; the greatest was obtained in the calibration performed according to the Stocking-Lord method when 5 anchors were used.

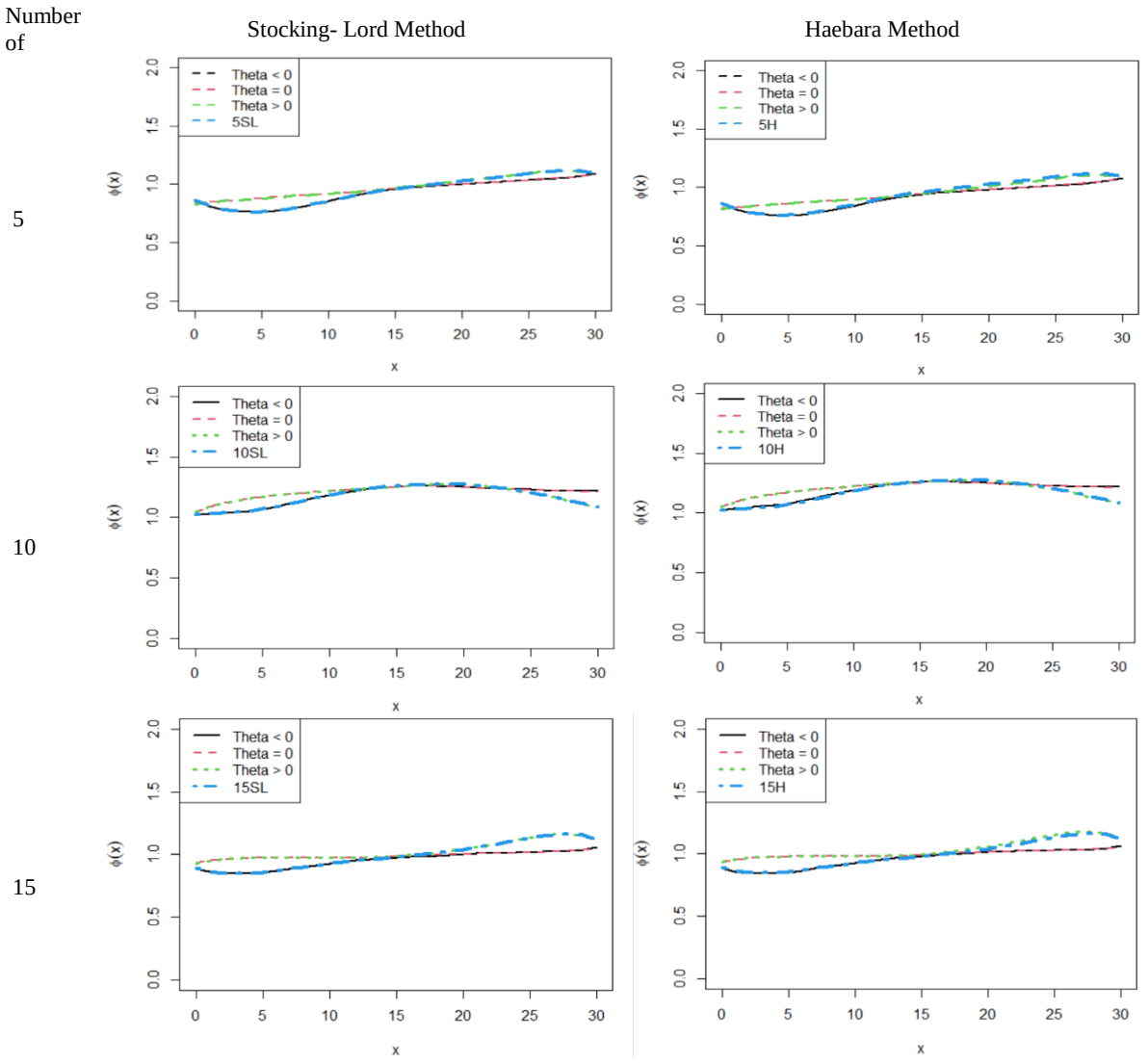

Figure 2. Equated Scores Obtained With Kernel Equating and Kernel Local Equating Based on IRT in Conditions Where the Number of Anchor Items are 5, 10 and 15, Respectively, and Function Graphs Regarding The Differences Of Raw Scores

The results of the Kernel local equating show that the errors of the equated scores were estimated .003 times higher with the Stocking-Lord method compared to the Haebara method. As the number of anchor items increased, equating errors decreased in both methods and both ability levels. In cases when 5, 10, and 15 anchor items were used, the smallest error was obtained with equalizations of the middle ability level. Similarly, in the equatings made according to low ability level, estimates were made with relatively high errors. In addition, in the equalizations made according to the middle ability level, errors are relatively more homogenous. The smallest of these errors was when the 15 anchor items were used according to the middle ability level with the Haebara method. The greatest error, on the other hand, was when 5 anchor items were used with the low ability level based on the Stocking-Lord method.

It was revealed that Kernel equating errors were greater than those of Kernel local equating when comparing Kernel equating and Kernel local equating in all conditions in all ability levels.

Table 3. Error Distributions Obtained From IRT Observed Score Kernel Equating and Kernel Local Equating

		Kernel Equating					Kernel Local Equating			
Calibration	Number of Anchor	Min.	Max.	Mean	S.D.	θ level	Min.	Max.	Mean	S.D.
Stocking-Lord	5	0.188	0.496	0.375	0.080	L	0.188	0.490	0.354	0.082
						M	0.189	0.367	0.312	0.047
						H	0.181	0.418	0.333	0.058
	10	0.147	0.365	0.278	0.056	L	0.144	0.362	0.263	0.060
						M	0.141	0.263	0.228	0.032
						H	0.139	0.302	0.243	0.040
	15	0.134	0.312	0.243	0.045	L	0.129	0.310	0.229	0.049
						M	0.127	0.232	0.201	0.027
						H	0.127	0.265	0.215	0.033
Haebara	5	0.186	0.491	0.371	0.079	L	0.187	0.482	0.350	0.080
						M	0.188	0.363	0.308	0.047
						H	0.180	0.418	0.330	0.059
	10	0.146	0.362	0.276	0.056	L	0.143	0.358	0.260	0.059
						M	0.140	0.261	0.225	0.032
						H	0.138	0.303	0.241	0.041
	15	0.134	0.309	0.240	0.045	L	0.129	0.305	0.226	0.048
						M	0.126	0.228	0.198	0.027
						H	0.127	0.264	0.213	0.034

Note. L: Low, M: Medium, H: High

DISCUSSION and CONCLUSION

In this study, the P and Q forms based on the 2PL model with different anchor item numbers (5, 10, 15) were evaluated for different ability levels [$\theta < 0$ (low), 0 (moderate) and $\theta > 0$ (high)]. Equating results of Stocking-Lord and Haebara methods were examined.

The present study used simulated data in which the anchor items were not included in the individual scores, unlike the studies of Öztürk-Gübeş and Kelecioğlu (2015), Pektaş and Kılınç (2016), Tanberkan-Suna (2018), in which the real data were used. Akın Arıkan (2017), used simulated data as well; however, she only compared the Haebara method in IRT true score equating and Kernel equating methods. Öztürk-Gübeş (2019), on the other hand, investigated the effect of multidimensionality on test equating and not included the change in the item numbers. Moreover, Wang et al. (2020) compared equapercentile equating, Kernel equating, and IRT Kernel equating methods.

Errors and function graphs were examined related to the difference between raw and equated scores in IRT observed score Kernel equating non-equivalent anchor test design when anchor items and calibration methods differ. The results revealed that there are differences and similarities between the equated scores, the distribution of the difference scores and errors in non-equivalent groups with anchor test design with Stocking-Lord and Haebara methods. Equated scores were estimated with a higher mean score when 5 anchor items were used in both calibration methods. In all the conditions, equated scores are lower than each score that can be obtained from the test. In cases when the anchor item numbers were the same, errors of the equated scores based on Haebara method were estimated lower. As the number of anchor items increased, the errors of the estimates in both methods were closer to one another. Wang et al. (2020) also obtained similar results where the number of items was 30 and 45 in the simulation. This finding is not supported by the findings of Uysal (2014), in which he found that error estimates with the Stocking- Lord method were lower than the Haebara method.

In addition, the present study investigated the functions and errors regarding the difference scores and equated scores when the item numbers and calibration methods differed. Both Stocking-Lord and

Haebara methods yielded similar results in cases that the same number of anchor items were used and the equatings of the scores and errors were conducted according to the low, middle, and high ability levels. The mean scores of low, middle, and high ability levels were the greatest with the 5 anchor items and the smallest with the 10 anchor items with both methods. In all conditions, the lowest mean score was obtained with the Haebara method and 10 anchor items according to middle ability level. The highest mean score, on the other hand, was obtained with the Haebara method and 5 anchor items according to low ability level.

When both methods are compared, mean scores obtained with Kernel local equating with 5 and 15 anchor items according to low ability level were estimated higher than Kernel equating. When 10 anchor items were used, the results of Kernel equating were the highest in all the conditions. Also, the graphs about the relationship between raw and equated scores showed that the range of difference scores were the narrowest when Kernel local equating were used regardless of the calibration method. The reason for this could be the fact that errors were estimated lower with the help of different equating functions based on the ability level and raw scores.

The lowest errors were estimated when both methods were used with 5, 10, and 15 anchor items. Moreover, errors were homogenous in the equatings based on middle ability level. The reason behind this result could be that the simulation data were simulated with normal distribution in the middle ability level ($b = 0$). This finding was supported by Wiberg et al. (2014), which suggests three different observed score Kernel local equating methods by combining local equating and Kernel equating and found that Kernel local equalization methods are quite stable against the changes in the accuracy and length of the anchor test in the non-equivalent groups anchor test design. The Kernel local equating errors were lower than Kernel equating errors when the two methods were compared. This finding is not supported by the results of the study of Wiberg et al. (2014) in which they found that the Kernel local equating method yielded higher standard errors than Kernel equating.

As a result, it was found that IRT observed score Kernel equating and Kernel local equating Stocking-Lord and Haebara methods can both be used and to keep the errors low, the number of anchor items should be kept higher. Also, Kernel local equating should be used with the ability level most appropriate to the ability distribution of the individuals. In future studies, different Kernel equating methods, different calibration types, and different data collection designs can be used to compare the observed score with the true score equating in cases where the anchor item is internal and external. Also, Kernel equating, and Kernel local equating methods can be examined using the equivalent groups design. In addition, equating errors can be examined by dividing ability levels in IRT Kernel local equating. The present study made use of the simulation data; a similar study can be conducted with real data set.

REFERENCES

Akın Arıkan, Ç. (2017). *Kernel eşitleme ve madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması* (Yayımlanmış Doktora Tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Andersson, B., & Wiberg, M. (2014). *IRT observed-score kernel equating with the R package kequate.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.372.8712&rep=rep1&type=pdf

Andersson, B., Bränberg, K., & Wiberg, M. (2020). *Package 'kequate'*. Retrieved from https://mran.microsoft.com/snapshot/2020-03-08/web/packages/kequate/kequate.pdf

Baker, F. B. (2016). *Madde tepki kuramının temelleri* [The basics of item response theory]. (N. Güler, Ed., & M. İlhan, Çev.). Ankara: Pegem Akademi. (1985)

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim K. H., Falk C. F., …, and Oguzhan, O. (2021). *Package 'mirt'*. Retrieved from https://cran.r-project.org/web/packages/mirt/mirt.pdf

Choi, S. I. (2009). *A comparison of kernel equating and traditional equipercentile equating methods and the parametric bootstrap methods for estimating Standard errors in equipercentile equating* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Diao, H. (2018). *Investigation repeater effects on small-sample equating: Include or exclude?* (Doctoral thesis). University of Massachusetts-Amherst.

Gök, B., & Kelecioğlu, H. (2014). Denk olmayan gruplarda ortak madde deseni kullanılarak madde tepki kuramına dayalı eşitleme yöntemlerinin karşılaştırılması. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 10*(1), 120-136. https://dergipark.org.tr/tr/download/article-file/161036 adresinden erişilmiştir.

González, J., & Wiberg, M. (2017). *Applying test equating methods: Using R*. Switzerland: Springer International Publishing. Retrieved from http://www.mat.uc.cl/~jorge.gonzalez/index_archivos/EquatingRbook.htm

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in highstakes testing. *Educational Measurement: Issues and Practice., 23*(1), 17-27. doi: 10.1111/J.1745-3992.2004.TB00149.X

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Baston: Kuluwer-Nijhoff Publisihing.

Holland, P. W., & Thayer, D. T. (1981). *Section pre-equating the graduate record examinations* (ETS Research Report Series). *1981*(2), i-62.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices.* New York: Springer.

Liou, M., Cheng, P. E., & Johnson, E. G. (1997). Standard errors of the Kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369. doi: 10.1177/01466216970214005

Norman Dvorak, R. L. (2009). *A comparison of kernel equating to the test characteristic curve method* (Unpublished doctoral dissertation). University of Nebraska-Linkoln.

Öztürk-Gübeş, N. (2019). Test eşitlemede çok boyutluluğun eş zamanlı ve ayrı kalibrasyona etkisi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 34*(4), 1061-1074. doi: 10.16986/HUJE.2019049186

Öztürk-Gübeş, N., & Kelecioğlu, H. (2015). Farklı test eşitleme yöntemlerinin eşitlik özelliği ölçütüne göre karşılaştırılması. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 48*(1), 299-214. doi: 10.1501/Egifak_0000001358

Pektaş, S., & Kılınç, M. (2016). PISA 2012 matematik testlerinden iki kitapçığın gözlenen puan eşitleme yöntemleri ile eşitlenmesi. *Mehmet Akif Ersoy Üniversitesi Eğitim Fakültesi Dergisi, 1*(40), 432-444. https://dergipark.org.tr/tr/download/article-file/264191 adresinden erişilmiştir.

Revelle, W. (2021). *Package 'psych'*. Retrieved from https://cran.rstudio.org/web/packages/psych/psych.pdf

Rizopoulos, D. (2018). *Package 'ltm'*. Retrieved from https://cran.r-project.org/web/packages/ltm/ltm.pdf

Tanberkan-Suna, H. (2018). *Grup değişmezliği özelliğinin farklı eşitleme yöntemlerinde eşitleme fonksiyonları üzerindeki etkisi* (Yayımlanmış Doktora Tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Uysal, İ. (2014). *Madde tepki kuramına dayalı test eşitleme yöntemlerinin karma modeller üzerinde karşılaştırılması* (Yayımlanmamış Yüksek Lisans Tezi). Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Enstitüsü, Bolu.

van der Linden, W. J. (2000). A test-theoretic approach to observed-score equating. *Psychometrika*, *65*(4), 437-456. Retrieved from https://link.springer.com/content/pdf/10.1007/BF02296337.pdf

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent groups design. *Journal of Educational and Behavioral Statistics, 33*(2), 186-203. doi: 10.3102/1076998607302633

von Davier, A. A. (2013). Observed-score equating: An overview. *Psychometrika*, *78*(4), 605-623. doi: 10.1007/s11336-013-9319-3

von Davier, A., Holland, P. W., & Thayer, D. T. (2004). *The Kernel method of equating.* New York: Springer.

Wang, S., Zhang, M., & You, S. (2020). A Comparison of IRT Observed Score Kernel Equating and Several Equating Methods. *Frontiers in psychology*, *11*, 308. doi: 10.3389/fpsyg.2020.00308

Wang, T., Lee, W. C., Brennan, R. J., & Kolen, M. J. (2008). A comparison of the frequency estimation and chained equipercentile methods under the common-item non-equivalent groups design. *Applied Psychological Measurement, 32*(8), 632-651. doi: 10.1177/0146621608314943

Wiberg, M., van der Linden, W. J., & von Davier, A. A. (2014). Local observed-score kernel equating. *Journal of Educational Measurement*, *51*(1), 57-74. doi: 10.1111/jedm.12034

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 358-373

Mixed Adaptive Multistage Testing: A New Approach

Anthony RABORN * Halil Ibrahim SARI **

Abstract
Computerized adaptive testing (CAT) and computerized multistage testing (CMT) are two popular versions of adaptive testing with their own strengths and weaknesses. This study proposes and investigates a combination of the two procedures designed to capture these strengths while minimizing the weaknesses by replacing the standard MST routing module with a CAT-based, item-level routing module. A total of 3000 examinees were simulated from a truncated normal distribution with bounds at -3 and 3, and a simulation study was conducted. Simulation results indicate that the new method provides some efficiency improvements over traditional MST when both routing modules are the same size, and when the item-level routing module is larger, the improvements are greater. The study showed that the proposed test administration model could be used to measure student ability, meaning that our new method resulted in lower mean bias, lower RMSE, and higher correlation than traditional MST. An R package built from the code used for this paper is also introduced in the supplementary file. The limitations of the study and recommendations for future research are also presented.

Key Words: Computerized adaptive test, multistage adaptive test, simulation, R, mixed adaptive test.

INTRODUCTION

There are two popular adaptive testing approaches: computerized adaptive testing (CAT) (Weiss & Kingsbury, 1984) and multistage testing (MST) (Luecht & Nungester, 2000). CAT is more widely known and more often used; in this approach, an examinee receives an item typically at medium difficulty level (e.g., maximizing information at the theta level of 0) and, based on his/her response to previous item(s), the item selection algorithm selects the next item from a large item pool. This continues until the examinee completes the test. A well-known advantage of CAT is allowing all test takers to work own personalized test producing high measurement accuracy in ability estimation (Yan, von Davier, & Lewis, 2016). In MST, however, the test has a panel design describing how different sets of items (e.g., 10 items) called modules are grouped into different stages. In stage one, there is typically one module called the routing module. In subsequent stages, there are several modules at different difficulty levels (e.g., easy, medium, and hard difficulty modules). An MST can be comprised of several stages and a different number of modules in each stage. For example, a 1-3-4 design has one module in stage one, three modules in stage two, and four modules in stage three. The working principle of MST is as follows. An examinee initially receives a set (e.g., 5 or 10 items) typically at the medium difficulty level. Based on the examinee's performance on this routing module, the module selection algorithm selects the next module from the next stage (Luecht, Brumfield, & Breithaupt, 2006). This continues until the examinee completes all the stages. The main difference between these two types of test administrations is that there is item-level adaptation in CAT but module-level adaptation in MST. Each has its own advantages and disadvantages.

MST has the disadvantage of being somewhat less efficient than CAT, meaning that CAT results in better theta estimates with lower standard errors than MST in many circumstances (Luecht & Sireci, 2012). This is due to item level adaptation feature of CAT. However, many common item-level adaptation schemes use maximum item information as the criterion for item selection, meaning the first few items selected by maximum information have higher exposure rates than later items; this can

* Supervisor, Accountability, Research, and Measurement, Pasco County Schools, Florida-United States of America, araborn@pasco.k12.fl.us, ORCID ID: 0000-0002-8083-4739
** Assoc. Prof., Kilis 7 Aralik University, Muallim Rifat Faculty of Education, Kilis-Turkey, hisari87@gmail.com, ORCID ID: 0000-0001-7506-9000

To cite this article:
Raborn, A., & Sarı, H. (2021). Mixed adaptive multistage testing: A new approach. *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 358-373. doi: 10.21031/epod.871014

Received: 30.01.2021
Accepted: 14.10.2021

be alleviated by modifying the item-level adaptation to choose from more than just the most informative item (Barrada, Olea, Ponsoda, & Abad, 2008). Another advantage of CAT over MST is that CAT allows for both varying and fixed test lengths, but the traditional MST is a fixed test length exam.

MST has the advantage of permitting test takers to answer or change answers to any question within the current module at any time while allowing for tests constructed to meet specific content and length requirements. This is an advantage because allowing response change provides having lower standard errors for the ability estimates, especially for students with higher abilities (Liu, Bridgeman, Gu, Xu, & Kong, 2015). Another advantage of MST is that it allows higher levels of control to test developers. This means that the developer of the test can place items into a module and easily keep track of content balancing, item usage, test length, or other statistical and non-statistical test requirements. However, these issues can sometimes be a problem in CAT, especially when there is a limited number of items from a content area in the item pool (Robin, Steffen, & Liang, 2016). Due to test assembly occurring prior to the test administration in MST, there is always greater expert control over item order and content area in this format (Sari & Huggins-Manley, 2017).

The better efficiency from CAT comes at the cost of the complex algorithms needed for the item-level adaptation, which MST avoids by having fewer adaptation points. This is because there are *n*-1 adaptation points in CAT, where *n* is the total test length as opposed to *k*-1 adaptation points in MST, where *k* is the number of stages. Having fewer adaptation points in MST has its own cost. For example, recovery of ability estimates becomes a difficulty when examinees are misrouted (e.g., incorrectly routed) through the modules. Previous research has shown that the initial routing stage has a major influence on the accuracy of final theta estimates, particularly in two-stage tests (Kim & Plake, 1993). Since the routing module provides the provisional theta estimate for the next modules, the routing module should include items from a wider range of difficulties. This means that it should maximize module level test information function at a wider theta range for test takers having different theta levels. Otherwise, it would be difficult to make better initial estimates for all test takers. A poorly designed routing module (e.g., with a very low maximum value for the test information function and/or very difficult or easy items) can place an examinee in the incorrect module in the subsequent stage. This would result in dramatic changes in the pathways one draws during the test. Consequently, it might be difficult or impossible to obtain less bias for the final theta estimate (Sari, Yahsi-Sari, & Huggins-Manley, 2016). As the number of stages increases, this influence is reduced, but practical considerations limit the number of stages that can be created and administered. Furthermore, previous studies showed that the reduction in estimation error provided by increasing the number of stages is modest (Patsula, 1999; Zenisky, Hambleton, & Leucht, 2010). A solution to establish better measurement accuracy after the routing module would be to increase the number of items in the routing module, but this would lead to an increase in the number of retired items after the test. This is because routing items are seen by all examinees and therefore reach maximum exposure rate.

Prior Attempts to Combine CAT and MST

A review of the literature showed that there was one other study that compared a proposed combination of CAT and MST. Wang, Lin, Chang, and Douglas (2016) performed three simulation studies investigating Hybrid Computerized Adaptive Testing, which used MST for the initial items and CAT for the subsequent items and compared it to traditional MST. Their hybrid test starts with MST (e.g., module-level adaptation) for the first two adaption points then uses CAT (e.g., item-level adaptation) for the remaining adaptation points. The first two simulations varied the proportion of items in the test that fell under the MST framework from $1/3^{rd}$ of the test length to $5/6^{th}$ of the test length and investigated six common MST designs, while the last simulation compared the two best designs from the first two simulations to two CAT and two MST designs. Their results indicated that, with two and three stages of various lengths, stage designs, and proportion of items in the MST stages, the hybrid designs (i.e., the combination of MST and CAT) perform as well or better than the traditional CAT design in terms of bias and RMSE and better than the studied MST designs in terms of RMSE.

In their study, the authors approached the problem primarily from the perspective of CAT (e.g., starting with MST and switching to CAT). In addition, their first simulation only used the two-stage 1-4 panel design for the MST comparison, and none of the three simulations fully compared the efficacy of the hybrid design to traditional MST designs. Thus, no single simulation included all of the factors manipulated in the study. This study, on the other hand, aims to follow the MST framework. Also, our study uses only three-stage designs but investigates the effect of MST design complexity and overall test length in the proposed hybrid design. The different emphases, as well as the different strengths of the approaches, lend credence to the investigation of ma-MST as an alternative to traditional MST and the other hybrid designs.

Purpose of the Study

In order to increase initial measurement accuracy while maintaining item exposure limits and allowing examinees to change answers within certain modules, we propose combining the CAT and MST methods into a single test administration. We called this new administration type as a mixed adaptive multistage test (ma-MST). The ma-MST will start with a CAT-based routing module (e.g., item-level adaptation) and obtain a provisional theta estimate. Then, this provisional theta estimate will be used to select the next MST-based stage. This means that the exam will start with CAT and switch to MST. We aim to bring MST closer to the efficiency of CAT while maintaining the aforementioned benefits of MST. By combining the methods in this way, the likelihood of misrouting can be reduced by the more accurate measure of ability after administrating items with item-level adaptation. As a result, this would result in a lower bias for the estimations of ability by the end of the test, while still allowing for easier control of item exposure rates, content balancing compared to the traditional MST, and allowing examinees the ability to change their answers in the later stages.

A New Approach: Mixed Adaptive Multistage Test (ma-MST)

Using R (R Core Team, 2016) and the R package "caMST" (Raborn, 2018), we investigated the efficacy of using item-level adaptation to route individuals to further modules. This new test format, mixed adaptive multistage test (ma-MST), is similar to a traditional MST in that it has a specific number of stages administered. However, the number of potentially administered tests is greater than in MST but less than in CAT because individuals would share panels of items depending on their ability estimates after seeing potentially different items in the CAT-based routing module.

This new method has much of the same test assembly processes as typical multistage tests and utilizes automated test assembly (ATA) to create each panel at each stage. In theory, ma-MST has similar item pool requirements as both CAT and MST. Item exposure concerns also remain and should be handled as appropriate for the use of the test (e.g., Reckase, 2010; van der Linden, 2000). In order to simplify the initial investigation of this method, there will not be any exploration of overall item exposure differences between CAT, MST, and ma-MST. This means that item exposure concerns will be ignored in favor of focusing on determining the accuracy of the different methods in their ability estimates.

In this study, the hybrid approach (e.g., ma-MST) will include a larger proportion of items selected with item-level adaptation points than in modules (e.g., resembling CAT). The ma-MST will also include a larger proportion of items in a module than selected with item-level adaptation points (e.g., resembling traditional MST). The primary goal of this study is to investigate the efficiency of the ma-MST, and what happens to the estimated theta parameters when the hybrid model resembles CAT and traditional MST. The expectation is that ma-MST would have lower bias and RMSE, higher correlation in the final theta estimations, especially as the CAT proportion increases.

For this study, we had two main research questions to answer:

1. How is the test efficiency (Bias, RMSE, and Correlation) be impacted when;
   a. CAT proportion (1/6, 1/3, and 2/3),

b. CAT item selection method (MFI, random selection),
c. MST designs (1-2-2, 1-2-3, and 1-3-3),
d. Test length (18 and 30 items) are varied in mixed approach simulations?

2. How will the test efficiency (Bias, RMSE, and Correlation) be impacted on the mixed adaptive and traditional MST under the combination of the levels of test length and MST design?

METHOD

We performed a simulation study to test the efficacy of the ma-MST against a traditional MST using the "caMST" package in R. The annotated R codes that demonstrate how to use the package to replicate the methods described here are provided in the supplementary file. We held constant the following factors: a) the number of stages (held at 3), b) the number of panels (3 parallel panels), c) the module selection or routing procedure (select the module with the maximum Fisher information [MFI] at the provisional theta), d) the initial ability estimate (held at $\theta_{\text{initial}} = 0$) and e) the provisional and final ability estimation procedures (expected a posterior [EAP], as commonly used in previous studies (Briethaupt & Hare, 2007; Luecht et al., 2006). The factors that we varied were the MST panel design, total test length, the fraction of the CAT routing module to the total test length, and the item selection procedure for CAT for a total of thirty-six conditions (see Table 1 for the levels). In addition to the ma-MST factors above, we used a traditional MST procedure as a baseline for each module design and test length.

Table 1. Simulation Study Conditions and Levels

Factor	Number of Levels	Levels
Panel Design	3	1-2-2 1-2-3 1-3-3
Test Length	2	18 items 30 items
CAT Module Length (fraction of overall test length)	3	1/6 1/3 2/3
Routing Module Item Selection	2	-Maximum Fisher information (Random 1 MFI) -Random selection from 5 items with Maximum Fisher information (Random 5 MFI)
	3x2x3x2=36	

The item parameters were based on a real Armed Services Vocational Aptitude Battery (ASVAB) military test used in Armstrong, Jones, Li, and Wu (1996). The simulated item bank had 450 multiple-choice items from four different content areas. In this study, in the 30-item condition, there were 10, 11, 4, and 5 items in content areas 1 through 4, respectively. For the 18-item condition, they are set to 6, 6, 3, and 3 items, respectively. The item parameters and the number of items for each content area in the original study were given in Table 2.

Table 2. Item Characteristics per Content Area

Content Area	*a*		*b*		*c*	
(Number of items)	Mean	SD	Mean	SD	Mean	SD
Content 1 (n = 150)	1.079	.409	-.467	1.179	.210	.095
Content 2 (n = 165)	1.128	.438	-.154	1.033	.200	.104
Content 3 (n = 60)	1.092	.538	-.025	.815	.203	.084
Content 4 (n = 75)	1.237	.383	-.014	.678	.162	.080

Armstrong et al. (1996)

A total of 3000 examinees were simulated from a truncated normal distribution with bounds at -3 and 3. Response patterns were generated according to Birnbaum's (1968) three-parameter (3PL) model in R. We used the EAP estimator (Bock & Mislevy, 1982) from the "mstR" package (Magis, Yan, & von Davier, 2017) with the prior distribution $N(0, 1)$ for all ability estimation. The IBM CPLEX program (ILOG, 2006) was used to construct the various modules in stages 2 and 3, and three essentially (although not strictly) parallel panels (i.e., the same number of items from the different content areas and similar in difficulty level). The items that were not used in these stages were treated as a mini item bank for the CAT and, depending on the test length and CAT proportion, the computer algorithm selected items from this bank consisting of the items remaining after the ATA. The bottom-up strategy was used when building the panels. The content distributions in the modules across the different test length and panel design conditions were given in Supplementary Tables 1, 2, and 3, under the 1/6, 1/3, and 2/3 CAT conditions, respectively. The panel-level test information across the CAT proportion and test length conditions were given in Supplementary Figure 1. For the modules in stages two and three, the module level information function was maximized at the fixed theta points of $\theta = -1$, $\theta = 0$, and $\theta = 1$ for the easy, medium, and hard modules in the conditions, respectively. In the baseline condition (e.g., traditional MST), the routing module was maximized at the theta point of 0.

Again, for the conditions with the CAT-based routing module, the items were selected from the pool of items that were not used for the modules. Then, for the random 1 MFI condition, the most informative item which fit the content area specification mentioned above was selected. For the random 5 MFI condition, a random item from the five most informative items which fit the content area specification was selected. This process was repeated after each item, updating the information function with every answer choice, until the simulated respondent answered the maximum number of items for the routing module.

The working principle of ma-MST simulation was as follows. In each design (e.g., 1-2-2, 1-2-3, or 1-3-3), if the CAT proportion was 1/6, and the total test length was 18, the computer tailored three items (1/6 of the 18 items) to the individual based on their responses in the first stage (e.g., item-level adaptation), and tailored 15 items in the two remaining stages (e.g., module-level adaptation). If the total test length was 30 and CAT proportion was 2/3, simulated individuals were administered 20 CAT-based items in the first stage and 10 total MST-based items in the second and third stages. This indicates that under the same total test length, as the CAT proportion increases, more items are administered at the item level.

To determine the efficiency of the tests within these conditions, we calculated mean bias, root mean squared error (RMSE), and Pearson correlations between true theta and estimated theta. It is important to note that each overall statistic was calculated for across the 3000 examinees for a replication (e.g., iteration) and averaged across 100 replications.

For the results, we ran a four-way factorial ANOVA separately for each of the outcomes, keeping the highest-order interaction terms in each case. To determine the magnitude of any experimental effects, the η^2 and partial η^2 statistics were calculated for each factor. Rather than using cut-off values for large effect sizes, the relative sizes of the η^2 statistics were compared within each outcome to determine which factor has the most influence on differences in the outcome measures. The findings of the simulation study are presented below.

RESULTS

Bias

The grand mean bias for each condition can be seen in Table 3. The largest bias (0.092) occurs within the 1-2-3 1/6 CAT 30-item MFI design, which appears larger relative to the other conditions. The smallest bias (.045) occurs within the 1-2-2 2/3 CAT 18 item random 5 MFI design. The smallest bias in the MST designs (.046) occurs within the 1-2-2 18 item design, while the largest bias in the MST

designs (.069) occurs within the 1-2-2 30-item design. Table 3 showcases the variability in bias and shows that the 1-2-3 design tends to perform the worst in the ma-MST designs.

Table 3. Grand Mean Bias Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	0.046	0.069	---	---
MST	1-2-3	0.054	0.061	---	---
MST	1-3-3	0.057	0.063	---	---
1/6 CAT	1-2-2	0.051	0.076	0.056	0.076
1/6 CAT	1-2-3	0.088	0.092	0.077	0.081
1/6 CAT	1-3-3	0.046	0.062	0.056	0.065
1/3 CAT	1-2-2	0.047	0.075	0.050	0.071
1/3 CAT	1-2-3	0.068	0.076	0.065	0.079
1/3 CAT	1-3-3	0.060	0.069	0.058	0.068
2/3 CAT	1-2-2	0.045	0.055	0.045	0.056
2/3 CAT	1-2-3	0.049	0.059	0.047	0.059
2/3 CAT	1-3-3	0.049	0.058	0.047	0.059

The ANOVA results for grand bias indicated that most interaction terms and main effects were significant (see Table 4), and the four-way interaction term remained in the model. However, the factors with the highest η^2 and η_p^2 were the main effects of test length (η^2 =.091, η_p^2 =.115) and CAT Proportion (η^2 = .089, η_p^2 =.112); these each explained about 11% of the unexplained variance in the mean bias. Panel design and the interaction between panel design and CAT proportion, the factors with the next largest η^2 and η_p^2, explained about 5% of the unexplained variance in the mean bias each. The other main effects, two-way and three-way interactions, were either non-significant or explained a very small proportion of mean bias variance.

Table 4. ANOVA Results for Grand Mean Bias

Factor	df	SS	MS	*F* value	*p*	η^2	η^2_p
Panel Design	2	2064	1032	367.85	.000*	.041	.055
Length	1	4566	4566	1627.25	.000*	.091	.115
CAT Proportion	3	4625	1542	549.38	.000*	.089	.112
Random	1	1	1	0.22	.641	.000	.000
Panel Design: Length	2	530	265	94.38	.000*	.011	.015
Panel Design: CAT Proportion	6	2149	358	127.64	.000*	.034	.046
Length: CAT Proportion	3	108	36	12.86	.000*	.002	.003
Panel Design: Random	2	72	36	12.84	.000*	.001	.002
Length: Random	1	11	11	4.09	.043	.000	.000
CAT Proportion: Random	2	11	6	1.97	.140	.000	.000
Panel Design: Length: CAT Proportion	6	310	52	18.43	.000*	.005	.007
Panel Design: Length: Random	2	10	5	1.73	.177	.000	.000
Panel Design: CAT Proportion: Random	4	157	39	13.99	.000*	.003	.004
Length: CAT Proportion: Random	2	89	45	15.91	.000*	.002	.003
Panel Design: Length: CAT Proportion: Random	4	45	11	4.05	.003*	.001	.001
Residuals	12558	35239	3				
Total	12599	49987					

* Significant at the .05 level.

RMSE

The grand mean RMSE for each condition can be seen in Table 5. The largest RMSE (0.339) occurs within the 1-2-3 1/6 CAT 18 item random 5 MFI design, while the smallest RMSE (0.225) occurs within the 1-2-3 2/3 CAT 18 item random 5 MFI design. For the MST designs, the largest RMSE (0.327) occurs within the 1-2-2 18 item design, while the smallest RMSE (0.269) occurs within the 1-3-3 30 İtem design.

Table 5. Grand Mean RMSE Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	0.327	0.277	---	---
MST	1-2-3	0.318	0.310	---	---
MST	1-3-3	0.280	0.269	---	---
1-6 CAT	1-2-2	0.312	0.319	0.286	0.289
1-6 CAT	1-2-3	0.339	0.337	0.299	0.296
1-6 CAT	1-3-3	0.328	0.331	0.286	0.289
1-3 CAT	1-2-2	0.299	0.307	0.264	0.269
1-3 CAT	1-2-3	0.301	0.307	0.264	0.271
1-3 CAT	1-3-3	0.308	0.309	0.266	0.271
2-3 CAT	1-2-2	0.270	0.278	0.226	0.238
2-3 CAT	1-2-3	0.268	0.278	0.225	0.239
2-3 CAT	1-3-3	0.270	0.280	0.229	0.240

The ANOVA results for grand mean RMSE indicated that the four-way interaction between the factors was not significant, so Table 6 shows the ANOVA without this interaction term. Two factors dominated the variance explained RMSE -test length and CAT proportion- despite the significance of most of the interaction terms and all the main effects. The test length explained 36.7% of the total variance in RMSE, while the CAT proportion explained 42.6% of the total variance in RMSE. No other factors or interactions explained more than 5% of the total or unexplained variance in RMSE.

Table 6. ANOVA Results for Grand Mean RMSE

Factor	df	SS	MS	F value	*p*	η^2	η^2_p
Panel Design	2	181	91	58.59	.000*	.001	.009
Length	1	46689	46689	30199.84	.000*	.367	.706
CAT Proportion	3	57206	19069	12334.12	.000*	.426	.736
Random	1	924	924	597.834	.000*	.007	.045
Panel Design: Length	2	59	29	19.07	.000*	.000	.003
Panel Design: CAT Proportion	6	1792	299	193.22	.000*	.008	.048
Length: CAT Proportion	3	154	51	33.17	.000*	.001	.005
Panel Design: Random	2	6	3	1.82	.162	.000	.000
Length: Random	1	1	1	0.57	.451	.000	.000
CAT Proportion: Random	2	307	153	99.28	.000*	.002	.016
Panel Design: Length: CAT Proportion	6	296	49	31.94	.000*	.002	.011
Panel Design: Length: Random	2	1	0	0.32	.724	.000	.000
Panel Design: CAT Proportion: Random	4	46	12	7.45	.000*	.000	.002
Length: CAT Proportion: Random	2	20	10	6.36	.002*	.000	.001
Residuals	12562	19421	2				
Total	12599	127103					

* Significant at the .05 level.

Based on Table 6, test length was the most important factor on the RMSE and followed by CAT proportion. As the test length or CAT proportion increased, the amount of RMSE decreased.

Correlation

The grand mean correlation between the true and estimated theta values for each condition can be seen in Table 7. The smallest correlation (0.949) occurs within the 1-2-3 1/6 CAT 30 item random 1 MFI design, while the largest correlation (0.980) occurs within the 1-2-2 2/3 CAT 30 item random 5 MFI design. The MST design with the smallest correlation (0.950) was the 1-2-2 18 item design, while the largest correlation (0.971) occurred in the 1-3-3 30 item design.

Table 7. Grand Mean Correlation Across Conditions

CAT Proportion	MST Design	Random 1		Random 5	
		18 Item	30 Item	18 Item	30 Item
MST	1-2-2	.950	.970	---	---
MST	1-2-3	.957	.957	---	---
MST	1-3-3	.970	.971	---	---
1-6 CAT	1-2-2	.955	.955	.968	.967
1-6 CAT	1-2-3	.951	.949	.966	.966
1-6 CAT	1-3-3	.954	.952	.970	.970
1-3 CAT	1-2-2	.960	.958	.974	.972
1-3 CAT	1-2-3	.959	.958	.973	.972
1-3 CAT	1-3-3	.958	.958	.971	.972
2-3 CAT	1-2-2	.967	.966	.979	.980
2-3 CAT	1-2-3	.967	.964	.979	.978
2-3 CAT	1-3-3	.967	.964	.978	.976

Note. All correlations were significant at the alpha level of .05

The ANOVA results for grand correlation can be seen in Table 8. Like the grand mean RMSE, the four-way interaction between all factors was not significant and was removed from the ANOVA. Additionally, the same pattern of η^2 and η_p^2 was found: the highest values were found for the test length and CAT proportion, which explain 62.2% and 24.4% of the total variance, respectively. No other factor or interaction of factors explained greater than 5% of the variance in mean correlations.

Table 8. ANOVA Results for Grand Mean Correlation

Factor	df	SS	MS	F value	*p*	η^2	η^2_p
Panel Design	2	1	1	6.92	.001	.000	.001
Length	1	6211	6211	84496.47	.000	.622	.871
CAT Proportion	3	2622	874	11891.51	.000	.244	.725
Random	1	32	32	440.50	.000	.003	.034
Panel Design: Length	2	0	0	2.07	.126	.000	.000
Panel Design: CAT Proportion	6	85	14	193.83	.000	.003	.034
Length: CAT Proportion	3	60	20	271.37	.000	.004	.042
Panel Design: Random	2	1	0	4.36	.013	.000	.001
Length: Random	1	1	1	20.33	.000	.000	.002
CAT Proportion: Random	2	7	3	44.94	.000	.001	.007
Panel Design: Length: CAT Proportion	6	44	7	99.49	.000	.002	.022
Panel Design: Length: Random	2	0	0	0.04	.956	.000	.000
Panel Design: CAT Proportion: Random	4	1	0	3.32	.010	.000	.001
Length: CAT Proportion: Random	2	0	0	1.60	.202	.000	.000
Residuals	12562	923	0				
Total	12599	9988					

Based on Table 8, test length was the most important factor on the correlation and followed by CAT proportion. As the test length or CAT proportion increased, the size of correlation decreased.

DISCUSSION and CONCLUSION

This study aimed to determine how useful ma-MST, which follows the MST framework but utilizes an item-level adaptation routing module as in CAT, is in estimating theta as compared to standard MST designs. We hypothesized that ma-MST performs better than MST under the current simulation conditions according to grand mean bias, grand mean RMSE, and grand mean correlation. The results indicated that replacing the routing module of a certain length in a traditional MST with an equal-length item level adaptation routing module as in CAT results in similar levels of bias, lower levels of RMSE, and higher levels of correlation between true and estimated theta value. Including even more CAT items at the initial stage (the 2/3 CAT conditions) resulted in somewhat larger improvements in bias, RMSE, and correlation. The best-case scenarios for each outcome measure occurred within a 2/3 CAT condition, while the worst-case scenarios occurred within a 1/6 CAT condition. The most likely explanation for these results is that in 1/6 CAT condition, there were fewer items administered with

item-level adaptation resulting in less accurate measures of ability in the routing stage, and the MFI item selection rule results in higher bias in the early stages of CAT (Chen, Ankenmann, & Chang, 2000).

The factors that were most important in determining the overall results were the test length and proportion of CAT items. Interestingly, tests with more items overall were associated with increased bias, although increasing the proportion of CAT items reduced the bias in every condition. This was counteracted with more items resulting in a smaller RMSE. This seeming contradiction is likely caused by a combination of the EAP estimator and by individuals at the boundaries of the module selection cutoffs (e.g., individuals with provisional ability estimates that caused the difference in the maximum module information to be small between the potential modules the individual could be routed to). The EAP estimator increases bias but decreases RMSE, particularly in more extreme values of ability (Kim, Moses, & Yoo, 2015). Improper routing of individuals is known to cause problems in MST, and the panel designs and module information functions in the simulation were not designed to prevent this from happening.

Unsurprisingly, we saw that the conditions with the highest proportion of CAT-based routing had the lowest levels of bias and RMSE as well as the highest correlations between the predicted and simulated theta values. However, since the ma-MST method provided good or better outcomes when the CAT routing panel was at least as large as a typical MST, the overall conclusion is that there is evidence to support the use of this design in circumstances that allow its use. For researchers and practitioners who wish to maintain many of the benefits of MST while improving its estimation efficiency, ma-MST is one method they should consider using.

While the study demonstrates the usefulness of ma-MST, it does so only for conditions that are similar to those in the simulation study. Another simulation with more varied conditions, such as different content balancing requirements, different unidimensional IRT models (e.g., 1PL or 2 PL), multidimensional IRT models, or estimation procedures, can further establish the usefulness of this approach, as well as a study comparing the designs with real data. Utilizing better panel designs which minimize the likelihood of misrouting or allow for misrouted individuals the chance to be re-routed into appropriate modules may provide more evidence of the efficacy of ma-MST over MST. Changes to the item and/or module selection method in ma-MST (e.g., by using a different information function) may also help improve the performance of the method as prior research has shown the choice of routing method can affect the efficacy of MST (Raborn, 2018). Another criticism in this study would be that the choice of some of the study conditions in the research design, especially for the ratio for the CAT proportion, is somewhat arbitrary. However, this study is an initial investigation of ma-MST approach. Finally, future research should investigate other ability estimation procedures such as maximum likelihood estimation as they may affect the relative efficiency of ma-MST when compared to MST.

As there have been other proposed combinations of CAT and MST in the literature such as Hybrid Computerized Adaptive Testing proposed by Wang et al. (2016), future research should include a comparison with these combinations as well as with full CAT tests. Investigating other simulation conditions that would serve to limit the limitations in this study would provide additional evidence for or against ma-MST in more circumstances.

REFERENCES

Armstrong, R. D., Jones, D. H., Li, X., & Wu, L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*(1), 89-98. doi: 10.1177/014662169602000108

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 493-513. doi: 10.1348/000711007X230937

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (ETS RR-81-20). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1981.tb01255.x

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, (Eds.), *Statistical theories of mental test scores* (pp. 17-20). Reading, MA: Addison-Wesley.

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

Briethaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement*, *67*(1), 5-20. doi: 10.1177/0013164406288162

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*(3), 241-255. doi: 10.1177/01466210022031705

ILOG. (2006). *ILOG CPLEX 10.0* [User's Manual]. Paris: ILOG SA. Retrieved from https://www.lix.polytechnique.fr/~liberti/teaching/xct/cplex/usrcplex.pdf

Kim, H., & Plake, B. S. (1993). *Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing.* Atlanta, GA: National Council on Measurement in Education.

Kim, S., Moses, T., & Yoo, H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing. *Journal of Educational Measurement*, *52*(1), 70-79. doi: 10.1111/jedm.12063

Liu, O. L., Bridgeman, B., Gu, L., Xu, J., & Kong, N. (2015). Investigation of response changes in the GRE revised general test. *Educational and Psychological Measurement*, *75*(6), 1002-1020. doi: 10.1177/0013164415573988

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden, & C. A. Glas, *Computerized adaptive testing: Theory and practice* (pp. 117-128). Netherlands: Springer.

Luecht, R. M., & Sireci, S. (2012). *A review of models for computer-based testing.* New York: The College Board. Retrieved from https://files.eric.ed.gov/fulltext/ED562580.pdf

Luecht, R. M., Brumfield, T., & Breithaupt, K. (2006). A testlet assembly design for adaptive multistage . *Applied Measurement in Education*, *19*(3), 189-202. doi: 10.1207/s15324818ame1903_2

Magis, D., Yan, D., & von Davier, A. (2017). *mstR: Procedures to generate patterns under multistage testing*. Retrieved from https://CRAN.R-project.org/package=mstR

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing* (Unpublished doctoral dissertation). University of Massachusett, Arherst, MA.

R Development Core Team . (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Raborn, A. W. (2018). *Package 'caMST'*. Retrieved from https://cran.r-project.org/web/packages/caMST/caMST.pdf

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological Test and Assessment Modeling*, *52*(2), 127-141. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1087.9450&rep=rep1&type=pdf

Robin, F., Steffen, M., & Liang, L. (2016). The multistage test implementation of the GRE revised general test. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing* (pp. 363-380). Boca Raton, FL: Chapman and Hall/CRC.

Sari, H. I., & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory & Practice*, *17*(5), 1759-1781. doi: 10.12738/estp.2017.5.0484

Sari, H. I., Yahsi-Sari, H., & Huggins-Manley, A. C. (2016). Computer adaptive multistage testing: Practical issues, challenges and principles. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 388-406. Retrieved from https://dergipark.org.tr/en/download/article-file/270019

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden, & G. A. W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 27-52). Dordrecht: Kluwer Academic Publishers. doi: 10.1007/0-306-47531-6

Wang, S., Lin, H., Chang, H. H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, 53(1), 45-62. doi: 10.1111/jedm.12100

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Yan, D., von Davier, A. A., & Lewis, C. (2016). *Computerized multistage testing: Theory and applications.* Boca Raton, FL: CRC Press.

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden, & C. A. Glass (Eds.), *Elements of adaptive testing* (pp. 355-372). New York: Springer.

Appendix A: Annotated R Codes

An early version of the caMST package was used to perform the analyses in this simulation study. The analysis could be performed in v0.1.0 of the package (available on CRAN and GitHub; a developmental version is also available on the first author's GitHub repository); this is the version used for the brief demonstration here. This walkthrough assumes that you have a working R installation have installed the package with 'install_packages("caMST")' or 'devtools::install_github("AnthonyRaborn/caMST")', and for simplicity's sake only two conditions are shown: the CMT condition with the 1-3-3 panel design, equal-length routing, stage 2, and stage 3 modules, and 18 items, and the Ma-MST condition with the 1-3-3 panel design, 1/3 CAT routing module, 18 items, and MFI item selection in the routing stage.

This version of the package can only handle dichotomous IRT models and requires that four item parameters be specified for each of the items as in the four-parameter logistic model (4PL; Burton & Lord, 1981). That means that item parameters, as in most computer adaptive tests, are treated as fixed, known quantities and when using models other than the 4PL the equivalent item parameters still need to be specified. For example, if the item parameters being used come from the Rasch model, the discrimination and upper asymptote parameters for each item should be set equal to 1 and the guessing parameter for each item set equal to 0. As our simulation used 3PL items, the upper asymptote for all of our items was equal to 1, but the other three parameters varied as described in the text.

The main functions for this analysis were the *multistage_test* function, used for traditional MST formats, and the *mixed_adaptive_test* function, used for the ma-MST format. The data used in this study were simulated as explained above; item parameters were saved in a data frame with items on the rows and item parameters on the columns. To use the item parameter data frame with either of these functions, it should have the item parameters in the following format: item discriminations in column 1 named *a*, item difficulties in column 2 named *b*, the pseudo-guessing parameter in column 3 named *c*, and the upper asymptote in column 4 named *u*. Additionally, column 5 should be used for identifying the content area in which each item should be placed (if content balancing is needed) and is named *content_ID*. As of now, the item parameters must be formatted in this way for the functions to work.

From here, the *multistage_test* function will be used to demonstrate how we used the package functions for this study, then we will return to the *mixed_adaptive_test* function to highlight the differences in how the Ma-MST method is used.

The main function arguments for the *multistage_test* function are as follows:

- mst_item_bank: a matrix or data frame with the items formatted as above that contains all of the items that are used within this test. The rows of this data frame may be named to allow for the responses to be matched to the correct items automatically.
- modules: a matrix that relates the items in mst_item_bank to the modules in which they belong
- transition_matrix: a matrix that describes the possible modules individuals may be routed through
- response_matrix: a matrix or data frame of individuals' responses to the items in mst_item_bank, with persons on the rows and items on the columns. The item responses may be in the same order as in mst_item_bank: the first column of response_matrix should be the item in the first row of mst_item_bank. If not, the columns should share the same naming format as the rows of the mst_item_bank data frame to allow for the responses to be matched to the correct items automatically.
- n_stages: a numeric value indicating the number of stages in the test (e.g., the number of adaptation points plus one for the routing stage).
- test_length: a numeric value indicating the total number of items individuals will see.

Other options exist which allow for greater control over the way the item responses are analyzed; the function documentation goes into more detail.

For the 1-3-3 18-item CMT condition, the items we used are included with the package and can be called with the following commands:

```
## library(caMST)
## data(mst_only_items)
```

This will create the mst_only_items object in your global environment, which is a data frame with 42 rows (items) and 5 columns (item parameters). Using the head() function on this object shows the first six items and their paramaters (see Table A1).

These items were already placed in order in terms of the module they came it; that is, since each module has six items and there are seven modules across the three stages, the first six items are in the routing module, the second set of six items are in the first module at the second stage (the easy module), the third set of six items are in the second moduel at the second stage (the medium module), and so on. The item-module matrix for this data can be called into the environment with

```
## data(mst_only_matrix)
```

and is a 42 row (items) by 7 column (modules) matrix that looks like in Equation A1.

$$\begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \quad \text{(A1)}$$

The next argument specifies the relationship between the modules (e.g., the lines in Figure 2). The 1-3-3 design we used in the simulation allows for individuals to move from one module in a stage to modules in the next stage that are the same difficulty or slightly more/less difficult, but does not allow for complete crossover. This means that a person routed to the stage 2 easy module may be placed into the stage 3 easy or medium difficulty modules but not the hard difficulty module. The transition matrix codifies this relationship using 0s to indicate that an individual in the row's module cannot be placed in the column's module and 1s to indicate that they could be placed from the row's module to the column's module. The matrix for this condition is called with

```
## data(example_transition_matrix)
```

and looks like in Equation A2.

$$\begin{matrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix} \quad \text{(A2)}$$

The transition matrix should always be a square matrix with row and column sizes equal to the number of modules in the data, and the rows for the final stage modules should always be filled with 0 because there is no transition after the test is complete!

The response_matrix is simply the matrix or data frame of person responses. The package functions will try to use the column names of the response_matrix and the row names of the mst_item_bank data frame to extract the responses relevant to the current condition. Since caMST can only handle binary items, the responses should be all 0s, 1s, and NAs. An example set of responses from this simulation can be called with data(example_responses), which populates the current R environment with a 5 row (individuals) by 600 column (items) data frame of responses.

With these objects, we can use the multistage_test function to analyze the item responses as a CMT with the following code:

```
## multistage_test(mst_item_bank = mst_only_items, modules = mst_only_matrix,
## transition_matrix = example_transition_matrix,
## response_matrix = example_responses, n_stages = 3, test_length = 18)
```

The function will output a list of the results, which includes two different estimates of the individuals' abilities, the standard error of measurement for each individual, a matrix of the final items seen by all individuals, a matrix of the final modules seen by all individuals, and a matrix of the responses the individuals made to the items that they saw.

By changing the mst_only_items, mst_only_matrix, example_transition_matrix, n_stages, and test_length, each of the conditions ran in the simulation can be tested. In addition, since we know the true theta values used in the simulation, the bias, RMSE, conditional bias, conditional RMSE, conditional SEM, and correlation between true and estimated values are easily calculated with functions that take the true and estimated theta values as arguments. If the above results were saved as an object called *CMT_results*, calling *CMT_results$final.theta.estimate.mstR* produces these estimates and could be used for one of the functions. For example, assuming the true theta values are saved as a numerical vector called *example_thetas*, you could run *cor(example_thetas, CMT_results$final.theta)* to estimate the correlation between the values the responses were simulated from and the estimates from the *multistage_test* function.

The Ma-MST conditions were run with the *mixed_adaptive_test* function, which follows the same principles as the CMT function. The major difference is that the *mixed_adaptive_test* function requires two item banks: one for the first stage with item-level adaptation (i.e., for the CAT-style routing module), and another for the second and third stages (i.e., the CMT-style stages). Additionally, the function allows for some control over the way the CAT adaptation in the first stage occurs.

The arguments specific to the CAT routing module are:

- cat_item_bank: the item bank formatted as described in the "multistage_test" function
- item_method: the method for choosing items in the first stage; defaults to "MFI" (Maximum Fisher Information), which we used in our simulation
- cat_length: how many items are seen in the first stage
- cbControl: a list used for content balancing (not used in this study)
- cbGroup: a factor vector used for content balancing (not used in this study)
- randomesque: an integer value. The item_method ranks items from best to worst; using MFI and randomesque=1, the most informative item based on the Fisher information and the current response pattern is chosen, while using MFI and randomesque=5 will randomly select one item from the five most informative items based on the Fisher information and the current response pattern.

The arguments specific to the CMT modules are:

- mst_item_bank: the item bank formatted as described in the "multistage_test" function
- transition_matrix: a matrix that describes the possible modules individuals may be routed through

When comparing the two functions, it is easy to see that the addition of the CAT items is the only real change in the function arguments. The following code calls the new objects (one for the routing module items, another for the second and third stage items) and runs the Ma-MST 1-3-3 design 18 items 1/3 CAT MFI condition:

```
## data(cat_items); data(mst_items)
## mixed_adaptive_test(cat_item_bank = cat _items, cat_length = 6, item_method = "MFI",
## randomesque = 1,
## mst_item_bank = mst_items, modules = mst_only_matrix,
## transition_matrix = example_transition_matrix,
```

response_matrix = example_responses, n_stages = 3)

The results for this function contain the same information as the previous function, but are in a list format where each individual's entire results are saved as one element of that list. This helps when keeping track of the items each individual saw: the function keeps a track of the item parameters of each item seen by each individual and provides the individualized test bank as a part of the output.

Since the results of this function are in a list, it takes a little more effort to use them to test how well the method performs in terms of person parameter recovery. The easiest way to do this is by using the *getElement* function, which takes an object and the name of the element you wish to extract, within the *sapply* function, which applies one function to each element of another object. Putting these together will extract the information into a vector, similar to what the *multistage_test* function outputs automatically. If the output of the *mixed_adaptive_test* function was saved as *results*, then running

sapply(results, getElement, "final.theta.estimate.mstR", simplify = T)

will output a vector of the final estimated theta values. This can then be used as explained above to investigate the efficiency of the Mca-MST method under the specific conditions used.

By modifying the various function arguments and the objects used in the functions, this study could be replicated or even expanded relatively easily. The package documentation includes other examples, as well as a function for performing fully CAT-formatted tests. The readme file and GitHub website provide somewhat more in-depth examples with visuals on the input and output data.

Table A1. The First Six Items for the CMT Condition

Item	*a*	*b*	*c*	*u*	content_ID
Item7	1.534	0.216	0.163	1.000	1
Item24	1.458	-0.136	0.070	1.000	1
Item165	1.696	-0.189	0.190	1.000	2
Item187	1.735	-0.024	0.097	1.000	2
Item303	1.410	0.243	0.068	1.000	3
Item458	1.446	-0.475	0.277	1.000	4

Note: a is the item discrimination, b is the item difficulty, c is the item pseudo-guessing parameter, and u is the upper asymptote of the item function.

Appendix B: Supplementary Tables and Figures

Table B1. Content Distributions in the Modules in the 1/6 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:2 C2:2 C3:2 C4:2	-	C1:2 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	-	C1:1 C2:2 C3:2 C4:2	C1:4 C2:3 C3:3 C4:3	-	C1:4 C2:3 C3:3 C4:3	C1:3 C2:3 C3:3 C4:3	-	C1:3 C2:3 C3:3 C4:3
Total	8	-	8	7	-	7	13		13	12	-	12
1-2-3	C1:2 C2:2 C3:2 C4:2	-	C1:2 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:4 C2:3 C3:3 C4:3	-	C1:4 C2:3 C3:3 C4:3	C1:3 C2:3 C3:3 C4:3	C1:3 C2:3 C3:3 C4:3	C1:3 C2:3 C3:3 C4:3
Total	8		8	7	7	7	13		13	12	12	12
1-3-3	C1:2 C2:2 C3:2 C4:2	C1:2 C2:2 C3:2 C4:2	C1:2 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:1 C2:2 C3:2 C4:2	C1:4 C2:3 C3:3 C4:3	C1:4 C2:3 C3:3 C4:3	C1:4 C2:3 C3:3 C4:3	C1:4 C2:3 C3:3 C4:3	C1:4 C2:3 C3:3 C4:3	C1:4 C2:3 C3:3 C4:3
Total	8	8	8	7	7	7	13	13	13	13	13	13

S = Stage, E = Easy, M = Medium, H = Hard module

Table B2. Content Distributions in the Modules in the 1/3 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:1 C2:1 C3:2 C4:2	-	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	-	C1:1 C2:1 C3:2 C4:2	C1:2 C2:2 C3:3 C4:3	-	C1:2 C2:2 C3:3 C4:3	C1:3 C2:3 C3:1 C4:3	-	C1:3 C2:3 C3:1 C4:3
Total	6		6	6		6	10		10	10		10
1-2-3	C1:1 C2:1 C3:2 C4:2	-	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:2 C2:2 C3:3 C4:3	-	C1:2 C2:2 C3:3 C4:3	C1:3 C2:3 C3:1 C4:3	C1:3 C2:3 C3:1 C4:3	C1:3 C2:3 C3:1 C4:3
Total	6		6	6	6	6	10		10	10	10	10
1-3-3	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:1 C2:1 C3:2 C4:2	C1:2 C2:2 C3:3 C4:3	C1:2 C2:2 C3:3 C4:3	C1:2 C2:2 C3:3 C4:3	C1:3 C2:3 C3:1 C4:3	C1:3 C2:3 C3:1 C4:3	C1:3 C2:3 C3:1 C4:3
Total	6	6	6	6	6	6	10	10	10	10	10	10

S = Stage, E = Easy, M = Medium, H = Hard module

Table B3. Content Distributions in the Modules in the 2/3 CAT Conditions Across the Different Designs

	18-item						30-item					
Design	S2E	S2M	S2H	S3E	S3M	S3H	S2E	S2M	S2H	S3E	S3M	S3H
1-2-2	C1:1 C2:0 C3:1 C4:1	-	C1:1 C2:0 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	-	C1:0 C2:1 C3:1 C4:1	C1:1 C2:1 C3:1 C4:2	-	C1:1 C2:1 C3:1 C4:2	C1:1 C2:1 C3:2 C4:1	-	C1:1 C2:1 C3:2 C4:1
Total	3		3	3		3	5		5	5		5
1-2-3	C1:1 C2:0 C3:1 C4:1	-	C1:1 C2:0 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:1 C2:1 C3:1 C4:2	-	C1:1 C2:1 C3:1 C4:2	C1:1 C2:1 C3:2 C4:1	C1:1 C2:1 C3:2 C4:1	C1:1 C2:1 C3:2 C4:1
Total	3		3	3	3	3	5		5	5	5	5
1-3-3	C1:1 C2:0 C3:1 C4:1	C1:1 C2:0 C3:1 C4:1	C1:1 C2:0 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:0 C2:1 C3:1 C4:1	C1:1 C2:1 C3:1 C4:2	C1:1 C2:1 C3:1 C4:2	C1:1 C2:1 C3:1 C4:2	C1:1 C2:1 C3:2 C4:1	C1:1 C2:1 C3:2 C4:1	C1:1 C2:1 C3:2 C4:1
Total	3	3	3	3	3	3	5	5	5	5	5	5

S = Stage, E = Easy, M = Medium, H = Hard module

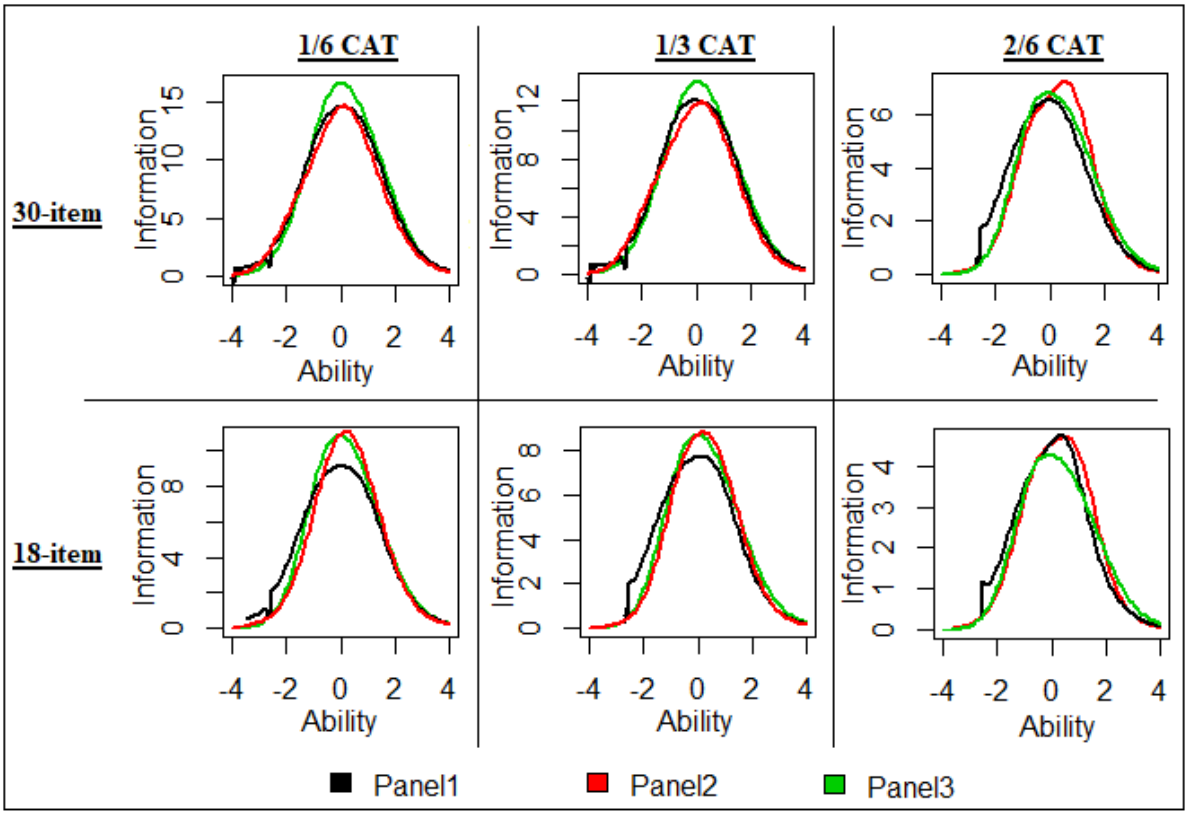

Figure B1. Plots for the Three Panels Under 1-2-2 ma-MST Design Across the 30-Item (Upper Three) and 18-Item (Down Three) and 1/6 CAT (Left two), 1/3 CAT (Middle two) and 2/6 CAT (Right two) Conditions

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(4); 374-387

Covariate Balance as a Quality Indicator for Propensity Score Analysis

Yusuf KARA* Akihito KAMATA** Elisa GALLEGOS***
Chalie PATARAPICHAYATHAM**** Cornelis J. POTGIETER******

Abstract
Propensity score analysis, such as propensity score matching and propensity score weighting, is becoming increasingly popular in educational research. When a propensity score analysis is conducted, examining the covariate balance is considered to be crucial to justify the quality of the analysis results. However, it has been pointed out that solely considering how covariates balance after matching may not be enough for justifying the quality of the propensity score analysis results. Suitable covariate balance may still yield biased estimates of treatment effects. The current study aimed to systematically demonstrate this problem by a series of simulation studies. As a result, it was revealed that a good covariate balance on the mean and/or the variance does not guarantee reduced bias on an estimated treatment effect. It was also found that estimation of the treatment effect can be unbiased to some degree, even with a lack of balance under specific conditions.

Key Words: Propensity score analysis, covariate balance, unbiased treatment effect.

INTRODUCTION

Propensity score (PS) analysis is becoming increasingly popular in educational research that adopts quasi-experimental design. PS analysis allows researchers to create a balance between treatment and control groups in order to estimate unbiased treatment effect when a randomized control design is not possible or is considered unethical/impractical (Guo & Fraser, 2015; Rosenbaum & Rubin, 1983). A challenge with the application of PS analysis is how to ensure that we have obtained an improved treatment effect estimate, ideally an unbiased estimate of the population treatment effect.

Typically, a researcher will go through a series of steps, and back steps, when conducting PS analysis. First, researchers identify the covariates to be included in the PS model and select a method (e.g., logistic regression-LR) for obtaining the propensity scores (PSs). Second, researchers conduct the analysis to estimate PSs and use them to balance the treatment and control groups in terms of covariate distributions. Third, researchers examine the quality of covariate balance and either go back to the first step and/or account for any insufficiently balanced covariates in the outcome analysis. Fourth, researchers conduct the outcome analysis for treatment effect estimation.

An important factor affecting PS analysis results is the type of covariates (i.e., associated only with the outcome, treatment assignment, or both) used to estimate PSs. Researchers have investigated the effects of including different covariate types in PS models in an effort to identify the most appropriate covariates

* Senior Data Analyst, Center on Research and Evaluation, Southern Methodist University, Dallas, TX, USA, ykara@smu.edu, ORCID ID: 0000-0003-0691-0630
** Professor, Department of Education Policy & Leadership and Department of Psychology, Southern Methodist University, Dallas, TX, USA, akamata@smu.edu, ORCID ID: 0000-0001-9570-1464
*** Senior Program Specialist, Center on Research and Evaluation, Southern Methodist University, Dallas, TX, USA, elisa@smu.edu
**** Research Assistant Professor, Department of Education Policy & Leadership, Southern Methodist University, Dallas, TX, USA, cpatarapichy@smu.edu
***** Assistant Professor, Department of Mathematics, Texas Christian University, Fort Worth, TX, USA, c.potgieter@tcu.edu, ORCID ID: 0000-0002-1995-6817

To cite this article:
Kara, Y., Kamata, A., Gallelogos, E., Patarapichayatham, C., & Potgieter, C.J. (2021). Covariate balance as a quality indicator for propensity score analysis . *Journal of Measurement and Evaluation in Education and Psychology, 12*(4), 374-387. doi: 10.21031/epod.993571

Received: 10.09.2021
Accepted: 19.12.2021

to include. Covariates that yield balanced treatment and control groups and ultimately unbiased treatment effects would be the desired ones to use in PS analysis. There have been some studies that revealed the importance of including covariates strongly associated with the outcome and not with the treatment assignment (Brookhart et al., 2006; Myers et al., 2011). Some studies also highlighted the negative effect of intrinsic covariates (mostly associated with the treatment assignment and showed little association with the outcome) on PS models and recommended avoiding them due to inconsistencies in the PS analyses (Bhattacharya & Vogt, 2007) or increased bias in the estimation of treatment effects (Patrick et al., 2011). Similarly, What Works Clearinghouse-WWC (2017) also highlighted the negative effects of using intrinsic covariates on the estimation of the treatment effect. Researchers also explored that including covariates those strongly associated with both treatment assignment and outcome, yields the least bias (Brookhart et al., 2006; Hong, Aaby, Siddique, & Stuart, 2018; Myers et al., 2011; Steiner, Cook, Shadish, & Clark, 2010).

Another important factor is the method utilized for obtaining PSs. The standard method is traditional LR in most applications. There have been more advanced methods that involve a combination of adjustments within the calculations/algorithms for obtaining PSs. For example, researchers have found that non-parametric and adaptive approaches (e.g., generalized boosted regression-GBR, classification and regression trees-CART, nearest neighbor matching-NNM, etc.) are more promising than LR in terms of covariate balance and treatment effect estimation (Cannas & Arpino, 2019; Lee, Lessler, & Stuart, 2010; McCaffrey, Ridgeway, & Morral, 2004; Setoguchi, Brookhart, Glynn, & Cook, 2008; Westreich, Lessler, & Funk, 2010). Other researchers found that certain PSs, such as the covariate balancing propensity scores (CBPS), perform very well in terms of balancing treatment and control groups (Kainz et al., 2017). In the current study, we consider multiple methods for estimating the PSs, which are elaborated in the methods section.

The most important overall goal for PS analysis is the reduction of confounding by balancing the treatment and control groups. In other words, PS analysis aims to reduce the confounding effect of external variables on the estimation of the true treatment effect. For this reason, examining covariate balance is crucial in justifying PS analysis results (Kainz et al., 2017). Researchers commonly evaluate the first moment of the covariate distributions between treatment and control groups by using the standardized absolute mean difference (SAMD). As also implied by Stuart, Lee, and Leacy (2013), SAMD is the most common measure of balance that is calculated similarly to effect size. It simply checks the magnitude of the mean differences in absolute scale compared to standardized mean difference (SMD). According to Rubin (2001), SAMD values from 0.1 to 0.25 are considered to be acceptable as indicators of good mean balance. Nevertheless, it was seen that applied researchers generally follow a stricter criterion as 0.05, which was suggested by WWC (2017) accessible through the Institute of Education Sciences as part of the US Department of Education.

In addition to checking the balance in terms of means, researchers may aim to evaluate other characteristics of covariate distributions in treatment and control groups. Along with SAMD, the literature has suggested the use of a combination of several criteria, including goodness-of-fit measures for covariate distributions (e.g., Kolmogorov-Smirnov test: Austin, 2009; Kainz et al., 2017; Stuart, 2010) and variance ratio (Kainz et al., 2017). Variance ratio (VR) is simply the ratio of a covariate's variance in the treatment and control group with a value of one indicating identical variances. According to Rubin (2007) VR values lower than 0.5 or higher than two are considered to be indicators of variance imbalance. We consider SAMD and VR as two widely-used standard measures of covariate balance in the current study. Readers are referred to Austin (2009) for a detailed overview of common balance measures and to Stuart, Lee, and Leacy (2013) for alternative balance measures such as prognostic score-based solutions.

Purpose of the Study

Obtaining good balance, such as a SAMD of 0.05 or less (What Works Clearinghouse, 2017), is standard practice when estimating treatment effects in PS analysis. Nevertheless, solely considering how covariates balance after matching may not be enough for justifying the quality of the PS analysis results.

Suitable covariate balance may still yield biased treatment effects (Lee, Lessler, & Stuart, 2010; Stuart, Lee, & Leacy, 2013). More specifically, Belitser et al. (2011) showed that following various balance diagnostic approaches might result in different levels of treatment effect bias. More interestingly, having measurement error in covariates might lead to a problematic estimate of the treatment effect even with a good level of covariate balance (Hong, Aaby, Siddique, & Stuart, 2018). Lastly, it was also shown that the balance of specific covariates could have more influence on the treatment effect bias (Stuart, Lee, & Leacy, 2013). All mentioned evidence from the literature points out the same conclusion: obtaining a good level of overall covariate balance might not be enough for estimating an unbiased treatment effect. There can be several other factors that can deteriorate the estimation of the true treatment effect even with a good amount of overall covariate balance.

Although there have been some studies that have discussed a potential lack of the direct relation between covariate balance and treatment effect bias (Hong, Aaby, Siddique, & Stuart, 2018; Lee, Lessler, & Stuart, 2010; Stuart, 2013), no study to our knowledge has demonstrated this problem by systematically examining it in the context of the aforementioned factors that are important to PS analysis, as well as other key factors such as sample size, the proportion of treatment group, and the association between covariates. Therefore, the current study aims to investigate the inconsistent relation between covariate balance and bias in treatment effect estimation. In other words, this study aims to systematically explore the effects of covariate balance on treatment effect estimation by considering many conditions that applied researchers encounter in their PS analyses. To facilitate a better examination of the current study results, we utilized the PSs estimated from different methods as weights for determining the treatment and control groups across all conditions (Guo & Fraser, 2015; Rosenbaum & Rubin, 1983).

METHOD

A simulation study was conducted to evaluate the performance of three PS estimation methods for recovering the population treatment effect and establishing the covariate balance in terms of means and variances. The three PS methods considered were the traditional LR, GBR, and CBPS. LR is the widely-used method among educational researchers and predicts the PSs through a logistic regression model. The GBR method uses a nonparametric, automated machine learning technique to estimate the PSs and associated weights (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017). The GBR method can predict the treatment assignment using a large number of covariates and is flexible in that it can handle nonlinear relationships between PSs and covariates (McCaffrey, Ridgeway, & Morral, 2004). The CBPS method simultaneously derives the PSs and weights for observations to optimize covariate balance between the treatment and control groups (Fong, Ratkovic, & Imai, 2019). Readers are referred to the cited literature for detailed explanations of the mentioned PS estimation methods.

Simulation Design

In addition to the three PS estimation methods, the current study also investigated the effect of using different types of covariates on the estimation quality of the treatment effect in relation to balance. Twenty-four covariates were classified into three different types (eight covariates per type) depending on their relationship with the treatment assignment and the outcome variable. Types of covariates were referred to as 1) type-W: correlated with both treatment assignment and the outcome variables, 2) type-X: correlated only with the outcome variable, and 3) type-Z: correlated only with the treatment assignment variable. A total of 108 simulation conditions were considered by crossing three PS estimation methods (LR, GBR, and CBPS), three covariate types (type-W, type-X, and type-Z), three sample sizes (500, 1,000, and 5,000), two proportions of treatment group (0.25 and 0.45), and two scenarios for the correlations among the covariates (uncorrelated and correlated).

Simulation conditions were mainly identified based on practical considerations that aim to guide applied researchers. LR was selected to represent the simplistic yet widely-used method among practitioners. CBPS and GBR were selected to represent more advanced methods compared to LR. It is known that CPBS is also a popular method in applied PS analysis studies and GBR is being recognized by many

practitioners who aim to use more advanced methods, namely machine learning-based approaches. Sample sizes were identified to reflect small to extremely large conditions. Group proportions were identified considering scenarios with low and medium/ levels of treatment group availability. It wouldn't be wrong to say that PS analysis studies mostly have larger sample sizes for the control group rather than the treatment group. Thus, we limited the maximum proportion of the treatment group to be 45% to represent a more realistic condition.

Some other magnitudes and/or parameters were fixed in the current simulation. The number of the covariates per type (fixed to eight) was identified randomly yet as a typical size of covariate availability in applied PS analysis studies. The magnitudes of the correlations between covariates were fixed to 0.05 and 0.15 for the uncorrelated vs. correlated covariates conditions. These values were identified in reference to the magnitudes used in other PS analysis simulation studies as well as thinking realistic magnitudes relative to the correlation values between covariates and treatment/outcome variables. We avoided high correlations among the covariates themselves in order to better reveal the effect of covariate and treatment/outcome relation. Lastly, the effect size was fixed to 0.8 and not varied in the current simulation. We chose a relatively high effect size in order to eliminate the side effects of having a small effect during the estimation phase. In other words, we intended to examine the performance of the methods under various conditions when the effect size is already known to be large.

Data Generation

Consider observations of the form $(\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i, B_i, Y_i),\ i = 1, \dots, n$ where $\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i$ are type-W, type-X, and type-Z covariates with size m_W, m_X, and m_Z respectively. As explained above, the number of the covariates for each covariate type was fixed to eight, thus $m_W = m_X = m_Z = 8$. Additionally, B_i is an indicator as to whether an observation belongs to the control ($B_i = 0$) or treatment ($B_i = 1$) group, and Y_i is the outcome of interest. As described earlier, it is assumed that covariates $\mathbf{W}_i$ and $\mathbf{Z}_i$ affect the probability of treatment group membership, while $\mathbf{W}_i$ and $\mathbf{X}_i$ affect the outcome after the group membership has been determined. For notational convenience, let $\mathbf{\Sigma}_W$, $\mathbf{\Sigma}_X$, and $\mathbf{\Sigma}_Z$ denote the covariance matrices of $\mathbf{W}$, $\mathbf{X}$, and $\mathbf{Z}$. Also, let $\mathbf{\Sigma}_{WX}$, $\mathbf{\Sigma}_{WZ}$, and $\mathbf{\Sigma}_{XZ}$ denote the cross-covariance matrices. Note that former matrices contain the covariances between the same-type covariates, whereas the latter ones contain the covariances between different covariate types. For example, $\mathbf{\Sigma}_W$ is the 8x8 covariance matrix of eight type-W covariates. $\mathbf{\Sigma}_{WZ}$ is the 8x8 covariance matrix of eight type-W covariates and eight type-Z covariates. Then $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_{WZ}$ matrices are combined for the generation of the treatment group membership. Other matrices can be interpreted in a similar way.

To simulate treatment group membership, let

$$\pi(\mathbf{W}, \mathbf{Z}) = \text{logit}(\alpha_0 + \boldsymbol{\alpha}_W^\top \mathbf{W} + \boldsymbol{\alpha}_Z^\top \mathbf{Z}). \quad (1)$$

Then, for the ith simulated case,

$$B_i \sim \text{Ber}[\pi(\mathbf{W}_i, \mathbf{Z}_i)]. \quad (2)$$

Note that π is the probability of being in the treatment group, and *Ber* stands for Bernoulli distribution. Also, terms with T superscripts refer to the transpose of a relevant matrix. The most important question here is how to choose the constant α_0 in (1), as this controls the proportion of cases that belong to the treatment and control groups. If $[\mathbf{W}_i, \mathbf{Z}_i]$ follows a zero-mean multivariate normal distribution (Φ is the inverse of the cumulative normal distribution function), then the choice

$$\alpha_0 = -\sigma_{WZ} \Phi^{-1}(1 - p) \quad (3)$$

with

$$\sigma_{WZ} = \sqrt{[\boldsymbol{\alpha}_W^\top, \boldsymbol{\alpha}_Z^\top] \begin{bmatrix} \boldsymbol{\Sigma}_W & \boldsymbol{\Sigma}_{WZ} \\ \boldsymbol{\Sigma}_{WZ} & \boldsymbol{\Sigma}_Z \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_W \\ \boldsymbol{\alpha}_Z \end{bmatrix}} \quad (4)$$

will result in a proportion p of the cases being associated with a success probability $\pi(\mathbf{W}, \mathbf{Z})$ greater than 0.5 and a proportion $1 - p$ with success probability less than 0.5. Once the treatment group memberships have been generated, the outcome Y_i is generated according to

$$Y_i = \beta_0 + \beta_T B_i + \boldsymbol{\beta}_W^\top \mathbf{W}_i + \boldsymbol{\beta}_X^\top \mathbf{X}_i + \varepsilon_i, \quad (5)$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ and $\beta_T = D.\sigma_\varepsilon$ with D being the effect size, which is fixed to 0.8, assuming the existence of a large treatment effect. The population values in (4) and (5) were set to $\boldsymbol{\alpha}_W = -1.0$, $\boldsymbol{\alpha}_Z = 1.0$, $\boldsymbol{\beta}_W = 1.0$, and $\boldsymbol{\beta}_X = 2.0$. Note that varying the population values of these parameters regulates the level of association between the covariates and treatment assignment/outcome. We selected these values somewhat arbitrarily yet in the light of the other PS simulation studies, including Brookhart et al. (2006).

The covariance matrices between the same- ($\boldsymbol{\Sigma}_W$, $\boldsymbol{\Sigma}_X$, and $\boldsymbol{\Sigma}_Z$) and cross-type covariates ($\boldsymbol{\Sigma}_{WX}$, $\boldsymbol{\Sigma}_{WZ}$, and $\boldsymbol{\Sigma}_{XZ}$) were varied depending on the magnitude of the relationship among the covariates as a simulation condition. Since all covariates were assumed to have a zero mean and a unit variance, covariance matrices were also the correlation matrices. For the conditions that assumed no relationship among the covariates, all correlations were fixed to zero. Thus, the same- and cross-type matrices were 8x8 identity and 8x8 zero matrices, respectively. For the conditions that assumed a relationship among the covariates, 0.15 and 0.05 were assigned to the correlations among the same- and cross-type covariates and were used to create the relevant covariance matrices.

Two hundred data sets were generated per simulation condition with R (R Core Team, 2018) and analyzed with the relevant PS analysis method, as elaborated in the next section. Note that the data were generated by considering all three types of covariates as predictors of the treatment assignment and outcome variables. In other words, the treatment assignment was generated by using the W- and Z-type covariates (8+8=16 predictors in total), and the outcome measure was generated by using the W- and X-type covariates (8+8=16 predictors in total). During the PS model fitting procedure, however, only one type of covariate (8 in total) was used for each analysis in order to examine the effect of covariate type as a simulation condition.

Analysis

PS weights were computed to estimate the average treatment effect for the treated (ATT), utilizing weighting by the odds (Hirano, Imbens, & Ridder, 2003). Therefore, weights for observations in the treatment group were fixed to be $w_{ti} = 1.0$ and weights for observations in the control group were computed by $w_{ci} = (ps_i)/(1 - ps_i)$, where ps_i is the estimated PS for the ith observation. w_{ti} and w_{ci} values were then used to compute weighted standardized treatment effect for the outcome variable. For covariates, the weights were applied to compute the SAMD and VR as the indicators of covariate balance per covariate. SAMD and the VR values (after adjusting for the treatment/control group as the denominator) for eight covariates were further averaged to obtain the overall balance indicators per simulated data set, which are referred to as average SAMD (ASAMD) and average VR (AVR) in the following sections.

We intended to use the default PS estimation options as much as possible for the three methods in their relevant R functions, considering a typical user without deep knowledge. Nevertheless, we modified some options for GBR method in order to prevent masking its performance compared to simpler methods. The traditional LR method was conducted by the *MatchIt* R package (Ho, Imai, King, & Stuart, 2008), adopting the default specifications for PS estimation with NNM. The GBR method was utilized by the *twang* R package (Ridgeway, McCaffrey, Morral, Griffin, & Burgette, 2017) with default specifications except for the stopping method, which was modified to assess the maximum balance matrix for Kolmogorov-Smirnov statistic. We also modified the estimand option to be ATT (as it was

the adopted method in the current study), which was ATE-Average Treatment Effect by default. The *CBPS* R package (Fong, Ratkovic, & Imai, 2019; Imai & Ratkovic, 2014) was used to estimate the PSs by the CBPS method with default options. Readers are referred to relevant resources for more details about the R functions of each method.

Evaluation Criteria

In an effort to derive a consistent measure of covariate balance and treatment effect, we utilized the *cobalt* R package (Greifer, 2019) to compute the ASAMD and AVR between treatment and control groups per generated data set. The *cobalt* package is commonly used as a supplement to the balance diagnostic tools and provides efficient summary tables of balance diagnostics for each covariate.

For the evaluation of overall covariate balance performance, ASAMD and AVR values were further averaged across 200 data sets generated per simulation condition. Average ASAMD values closer to 0 and average AVR values closer to 1 are considered to be good indicators of overall covariate balance in terms of mean and variance, respectively. Recovery of the population treatment effect ($D = 0.8$) was evaluated by the absolute bias (AB) and standard error (SE), which were calculated by equations (6) and (7), respectively.

$$AB(\widehat{D}) = \left|\overline{\widehat{D_m}} - D\right| \tag{6}$$

$$SE(\widehat{D}) = \sqrt{\frac{\sum_m^M (\widehat{D} - \overline{\widehat{D_m}})^2}{M}} \tag{7}$$

Note that D and $\widehat{D}$ are the population and estimated values of the treatment effect. Also, M represents the total number of replications (200 in our case), m is a specific step of those M replications, and $\overline{\widehat{D_m}}$ is the average of the estimated effect sizes across M replications.

RESULTS

Recovery of the Treatment Effect

It was confirmed that the use of type-W covariates provided the lowest ABs for the recovery of the treatment effect under all simulation conditions (Figure 1). This result is consistent with the existing literature that encourages researchers/practitioners to use covariates that correlate with both outcome and treatment indicators (Brookhart et al., 2006; Hong, Aaby, Siddique, & Stuart, 2018; Myers et al., 2011; Steiner, Cook, Shadish, & Clark, 2010). However, contrary to our expectation, the performance of type-X covariates (correlated with the outcome) was not as good as type-W covariates. Rather, their performance was close to type-Z covariates (correlated with the treatment condition), especially when there was no correlation among covariates. Nevertheless, the ABs were lower for type-X covariates (all below 0.6) than type-Z covariates for all simulation conditions. Lastly, the bias was slightly lower when the type-W covariates were correlated vs. they were not.

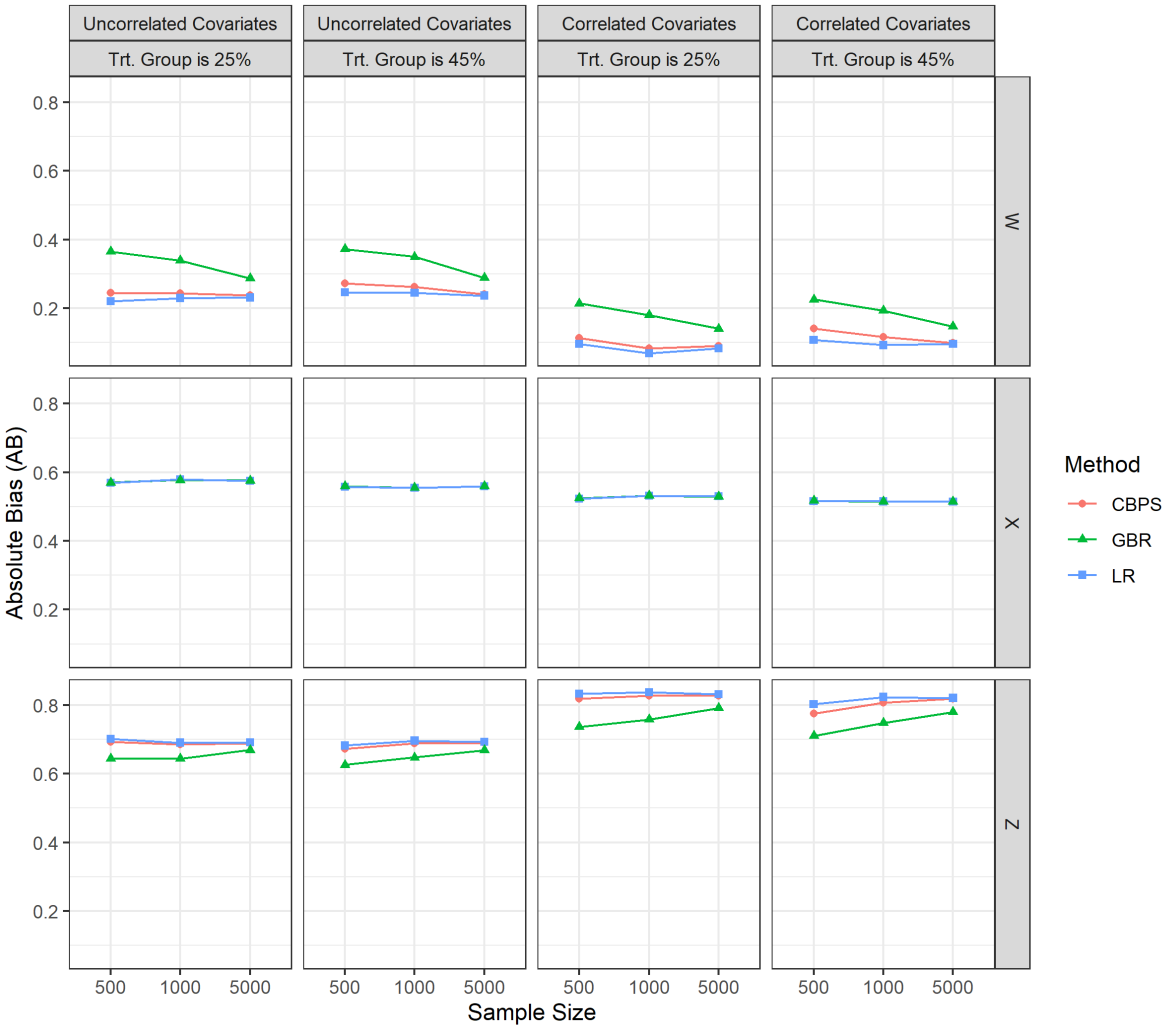

Figure 1. Absolute Bias of the Estimated Treatment Effect

In general, the three PS estimation methods performed similarly for recovering the population treatment effect under common conditions. This was especially true for conditions with type-X covariates. However, for conditions with type-W covariates, GBR did not perform as well as CBPS and LR, while CBPS and LR performed similarly. These tendencies were observed for all conditions, regardless of sample size, proportion of the treatment group, and correlation among covariates. Interestingly, GBR had the lowest AB values when type-Z covariates were used. However, the ABs were large, and the differences between GBR and the other two methods were not large enough to claim that GBR would be useful with type-Z covariates.

Regarding the sample size, there was no clear effect on the AB values for any of the three PS estimation methods. Only a slight decreasing tendency of AB with increased sample size was observed with the use of GBR method in conditions with type-W covariates. On the other hand, a clear effect of sample size was observed for SE values (Figure 2), which were in decreasing trend with the increase of sample size as expected. The proportion of the treatment group also did not show a considerable effect for AB values. Lastly, the effect of the correlation among the selected covariates showed different trends depending on a specific condition. For example, the AB values were lower when type-W covariates were correlated regardless of the proportion of the treatment group. This was also the case for type-X covariates; however, the change was not as clear as for the type-W covariates. An opposite tendency was observed for type-Z covariates. Namely, the AB values were higher when the covariates were correlated for both proportions of the treatment group. Thus, the amount of correlation between type-W

covariates seems to provide extra information for a better recovery of the treatment effect. Conversely, intercorrelated type-Z covariates seem to deteriorate the estimation of the treatment effect. Although we don't have an exact explanation for this phenomenon, we suspect that the intercorrelated treatment assignment predictors led to problematic balancing hence to slightly higher bias in the treatment effect estimation. Moreover, it is known that the use of only type-Z predictors is expected to result in a higher bias in the treatment effect estimation (Patrick et al., 2011). Thus, with the intercorrelated type-Z covariates, this negative effect seemed to be slightly larger.

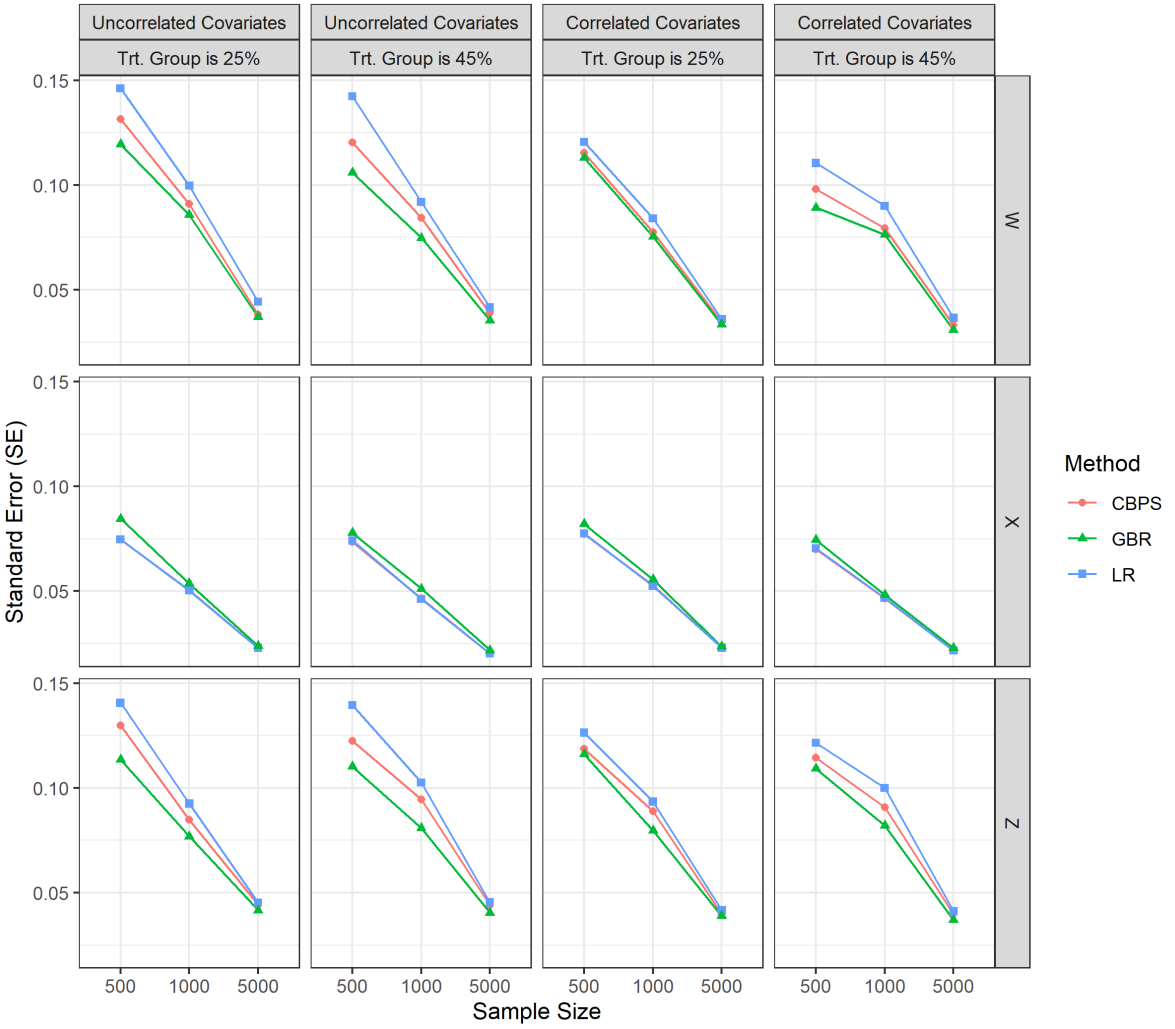

Figure 2. Standard Error of the Estimated Treatment Effect

In summary, our results demonstrated that the use of covariates that are related to both treatment indicator and the outcome resulted in a better recovery of the treatment effect nearly under all studied simulation conditions. Also, it was demonstrated that LR and CBPS produced better performance than GBR when type-W covariates were used. It is important to note that such a result might be explained by the data generation model, where we used only the first-level terms in the logistic regression of the treatment assignment. In other words, no higher-level terms (such as square of the predictors) or interactions were used in the generation of the treatment assignment. Thus, this is in line with the estimation of a plain logistic regression model adopted in LR method. GBR is known to examine also the prediction power of higher-level and interaction terms automatically. Thus, better performance of the LR compared to a more advanced method like GBR might be a result of this. Additionally, if selected

covariates are correlated only with the outcome, selection of the PS estimation method would not matter so much, as the three performed similarly.

Covariate Balance

Means

It was demonstrated that the mean balance was consistently better for the type-X covariates (correlated only with the outcome) under all conditions (Figure 3). Furthermore, the average ASAMD values for the type-X covariates were always below 0.05 regardless of the simulation condition, indicating that they met the threshold by WWC- What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education (2017). Therefore, a practitioner would interpret that the PS analysis went well. This result requires special attention, as it was demonstrated in the previous section that type-X covariates did not produce good AB for the recovery of the population treatment effect, compared to type-W covariates (correlated with the outcome and treatment condition). In other words, these results demonstrate that good covariate balance on means does not necessarily guarantee a good estimate of the treatment effect.

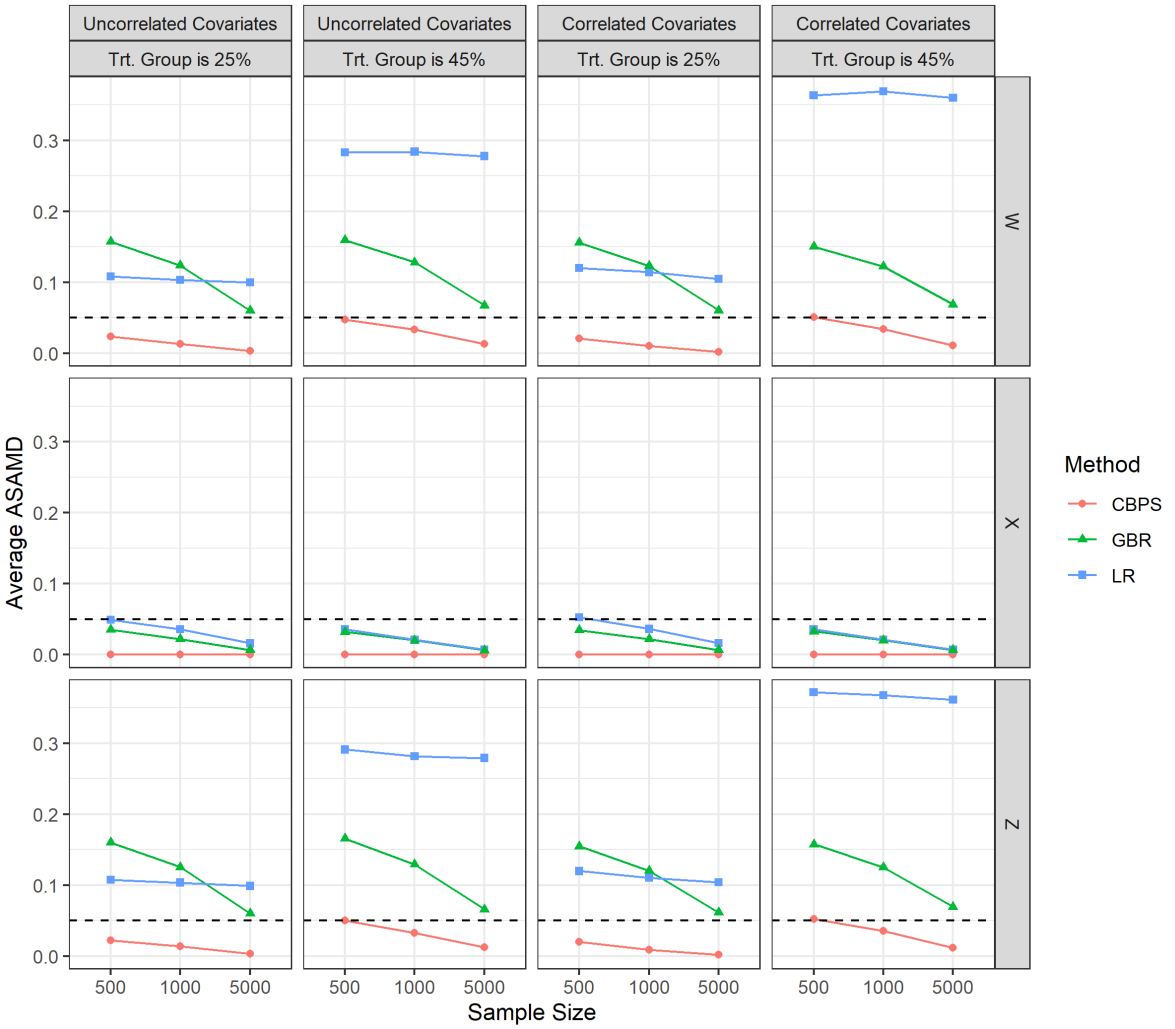

Figure 3. Covariate Balance on Means

On the other hand, when type-W covariates were used, the average ASAMD values met the WWC criterion of 0.05 or lower only in conditions with CBPS as the method used for PS analysis. It was also observed that conditions with GBR approached the WWC criterion with the largest sample size (n = 5,000) by the use of type-W covariates. Nevertheless, none of the two conditions resulted in low AB values, and GBR's AB values were the highest with the use of type-W covariates. Thus, these findings also support the fact that good mean-based covariate balance does not necessarily result in a less biased treatment effect estimate.

Looking at the results with the use of type-Z covariates (correlated only with the treatment condition) a similar trend was observed, where CPBS had the uniformly low average ASAMD values and GBR approached the WWC's criterion with the highest sample size. Nevertheless, the lowest AB was observed consistently for GBR, and surprisingly, the AB values were increased by the increase of the sample size. It is also worth noting that the average ASAMD values for LR were severely affected by the proportion of the treatment group when type-W and type-Z covariates were used. Related to this, LR was not the best performing method in terms of mean balance, compared to its good performance in terms of recovery of the treatment effect. This difference is more important for the use of type-W covariates. A potential explanation for this might be the simplicity of the LR method for computation of the PSs compared to CBPS and GBR. In other words, simple LR does not account for complex relationships during the balancing procedures as CBPS and GBR do.

Variances

Overall, GBR outperformed the other two PS estimation methods, producing better covariate balance with respect to variance ratio. This is not surprising because we set up GBR to derive PSs by evaluating the maximum of the balance matrices based on Kolmogorov-Smirnov statistic, which evaluates the difference in distribution shapes, as opposed to the difference in the means only. Exceptions were observed when the treatment group was 45% and n = 5,000 for type-W (correlated with the outcome and treatment condition) and type-Z covariates (correlated only with the treatment condition), where CBPS produced slightly better variance ratios than GBR. Unlike the mean balance of covariates, it was clear that the variance balance was affected by the sample size (Figure 4). All three PS estimation methods provided better variance ratios in conditions with larger sample sizes. It was demonstrated that sample size mattered more for CBPS than the other two methods.

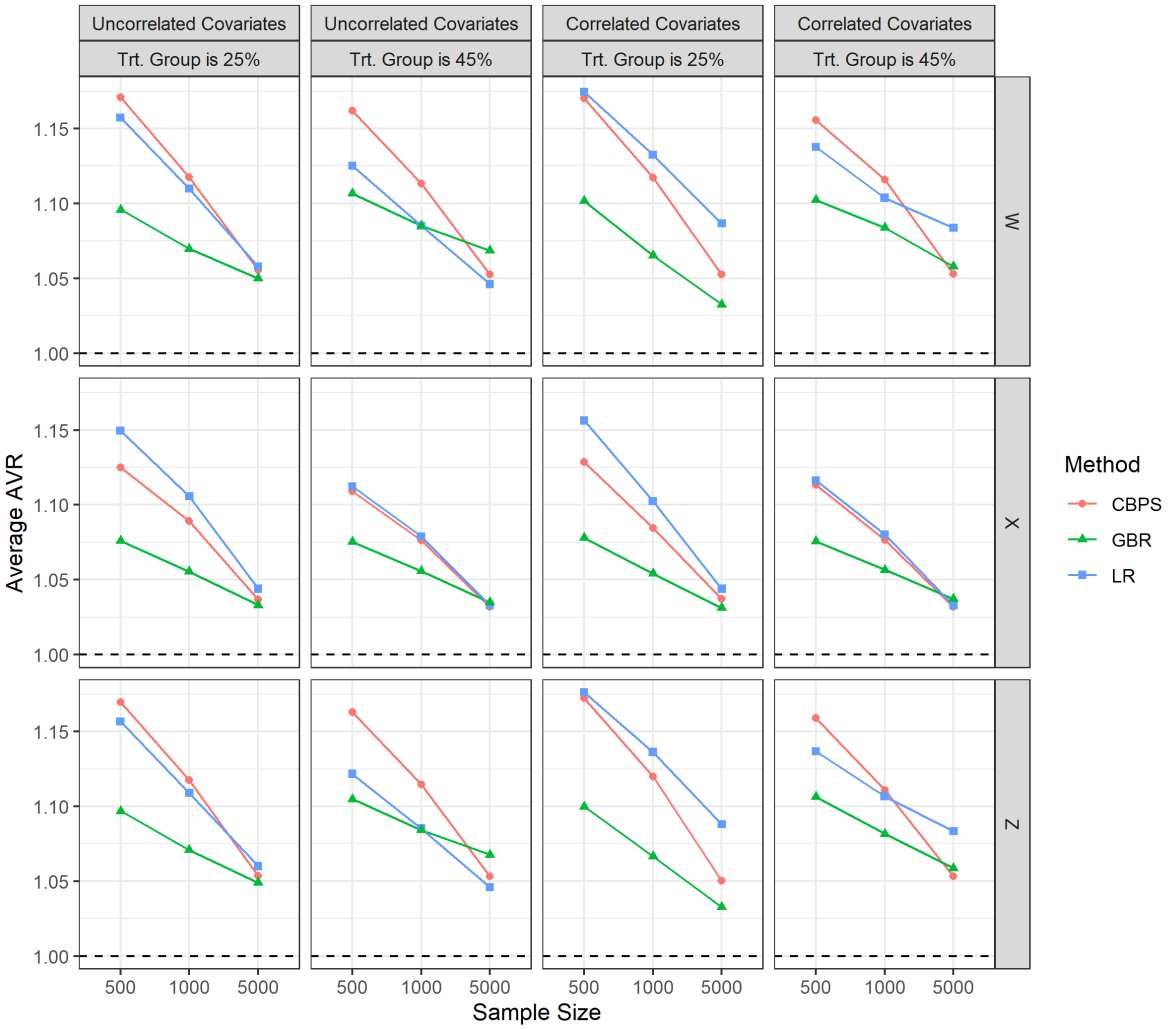

Figure 4. Covariate Balance on Variances

Similar to mean balance results, it was revealed that a PS estimation method that provided good variance balance did not necessarily do the same for the estimation of the treatment effect. Based on this determination, we can summarize our findings under two main statements. First, similar to mean balance results, the variance ratios were generally better for the type-X covariates (correlated only with the outcome) than the type-W and type-Z covariates across all conditions. However, as pointed out in the previous section, type-X covariates did not produce better AB values for the recovery of the treatment effect. Second, although GBR generally provided better variance ratios compared to the other two methods, it did not perform as well, in terms of AB, as CBPS and LR under conditions with type-W covariates. Also, LR provided the worst variance ratios for all conditions with type-X covariates; however, in terms of recovery of the treatment effect, LR performed as well as CBPS and GBR in conditions with type-X covariates. Therefore, based on our simulation results, we cannot conclude that variance ratios provided additional information to identify less biased treatment effect. Nevertheless, they provide more information about covariate balance in addition to mean balance.

DISCUSSION and CONCLUSION

Based on the simulation results, obtaining good covariate balance in terms of mean and variance ratio is likely when covariates are correlated only with the outcome variable (i.e., type-X covariates). The average mean balance across all conditions for this covariate type was below 0.05 on the standardized

scale, which meets the WWC criterion. The variance ratio for type-X covariates was consistently lower than other types of covariates in comparable conditions. However, the recovery of the treatment effect was not the best when only this type of covariates was used. Therefore, researchers/practitioners need to be cautious while using covariates that are mainly correlated with the outcome and not with the treatment assignment. Moreover, it is clear that in applied research that utilizes PS analysis, the true treatment effect will never be known. Thus, practitioners are encouraged to pay as much attention to the characteristics of the covariates as the level of balance they obtain after their selection. If the availability of the covariates is somewhat limited, practitioners can rely on the strengths of other PS estimation methods. For example, it was demonstrated that GBR provided slightly lower AB values when the covariates were the ones that only related to the treatment assignment. Nevertheless, we do not recommend solely relying on such an improvement since the bias values still were high in absolute values.

Also, obtaining a good mean covariate balance is more likely when the CBPS method is used regardless of the covariate type used, whether covariates are correlated or not, sample size, and the proportion of the treatment group. In all of the conditions this study investigated, the mean balance met the WWC criterion with the CBPS. However, it was revealed that the magnitude of AB was affected more by the types of covariates used, as already implied above. When the CBPS was used with covariates that are correlated with the outcome variable and treatment condition (i.e., type-W covariates), AB was quite small, especially when covariates were correlated with each other. However, it is not only for the CBPS; LR performed equally well, and the performance of the GBR was just slightly worse than the CBPS and LR.

Can an examination of the second moment (i.e., the variance ratio) help researchers/practitioners evaluate/predict the quality of an estimated treatment effect? Not likely. When the sample size was large, the variance ratio became very close to 1.05 for CBPS and GBR. However, this happened for all covariate types, including the ones that are correlated only with the treatment assignment (i.e., type-Z covariates). On the other hand, when the sample size was small ($n = 500$), variance ratios for CBPS and LR were large, up to 1.20 in the condition with correlated covariates with a 25% treatment group ratio. In conclusion, we can't recommend that using VR as a complement to ASAMD will be strong enough to predict the performance of the PS methods that would lead to less bias of treatment effect estimation. Rather, practitioners should pay more attention to the characteristics of the covariates they are planning to use for the estimation of PSs rather than solely relying on the level of balance.

It would not be wrong to say that CBPS can be suggested as the optimal method considering the general simulation conditions since it showed the best performance for the mean balance and better or nearly equal performance with two other methods for the recovery of the treatment effect. On the other hand, practitioners who mainly use type-X covariates would feel better by the good mean balance and variance ratio diagnostics they get. Nevertheless, they should be cautious about the estimation of the treatment effect since the type-W covariates (correlated both with the outcome and the treatment condition) showed better recovery results. In conclusion, it can be suggested to use covariates that are equally relevant to the treatment assignment and the outcome.

Limitations and Future Research

As explained in the methods section, specifications of the population values for the coefficients of different covariate types were somewhat arbitrary. It is likely that the results may change based on different specifications of those population values during the data generation phase. This is true for changing either the LR coefficients for the generation of the treatment indicator or the linear model for the generation of the outcome variable. Nevertheless, we tried to assign reasonable values for those parameters depending on their relation with the treatment condition and the outcome variable. We also checked other studies (e.g., Brookhart et al., 2006) as references for identifying typical values that are expected to be encountered in applied research. Also, data generation models did not assume any higher-level terms (e.g., quadratic effects) or interactions between covariates.

The number of covariates were limited to eight in derivations of PSs. Although this number is realistic in many applications of PS analysis, a larger number of covariates may change the results. Also, our simulations investigated three different covariate types (i.e., type-W, type-X, and type-Z) in turn only, meaning none of our PSs were estimated using a combination of different covariate types. Although this does not sound realistic, we intentionally performed that in order to reveal the isolated effect of each covariate type. This also mimics the scenario where applied researchers miss using a specific type of covariates that potentially might change the results of the PS analyses.

Although our study indicates that practitioners utilizing PS analysis should not rely on mean-based covariate balance, it is still unclear which diagnostic measure is ideal when conducting PS analysis. It could be that the ideal diagnostic measure depends on the method used to estimate the PSs (e.g., LR, GBR, CBPS) or on the approach used to apply the PSs (e.g., weighting, subclassification, etc.) in balancing treatment and control groups. The *cobalt* R package is able to work with the PS methods we explored in this study to provide weights. It could be that the "power" behind each method relies on using the weighting values generated from within each method rather than pulling the PSs to generate weights outside of each method. Future studies could investigate how different PS methods in combination with the *cobalt* R package generate weights and how these might affect covariate balance diagnostics and treatment effect bias.

Last, this study systematically demonstrated the effect of various conditions on covariate balance and estimation of the treatment effect through a series of simulated data. While results were quite promising, an empirical data set was not analyzed. We believe that a real data set would be helpful to confirm our simulation findings. Therefore, an empirical data analysis can be considered in a future study by using various PS estimation methods.

REFERENCES

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Belitser, S. V., Martens, E. P., Pestman, W. R., Groenwold, R. H. H., Boer, A. de, & Klungel, E. H. (2011). Measuring balance and model selection in propensity score methods. *Pharmacoepidemiology and Drug Safety*, *20*(11), 1115–1129. https://doi.org/10.1002/pds.2188

Bhattacharya, J., & Vogt, W. B. (2007). Do instrumental variables belong in propensity scores? Cambridge, MA: National Bureau of Economic Research (NBER) Working Paper Series No. 343.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, *163*(12), 1149–1156. https://doi.org/10.1093/aje/kwj149

Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. *Biometrical Journal*, *61*(4), 1049–1072. https://doi.org/10.1002/bimj.201800132

Fong, C., Ratkovic, M., & Imai, K. (2019). *CBPS: Covariate balancing propensity score*. Retrieved from https://CRAN.R-project.org/package=CBPS

Greifer, N. (2019). *Cobalt: Covariate balance tables and plots*. Retrieved from https://CRAN.R-project.org/package=cobalt

Guo, S., & Fraser, M. W. (2015). *Propensity score analysis: Statistical methods and applications (ed. 2)*. Thousand Oaks, CA: Sage.

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*(4), 1161–1189. https://doi.org/10.1111/1468-0262.00442

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2008). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Hong, H., Aaby, D. A., Siddique, J., & Stuart, E. A. (2018). Propensity score-based estimators with multiple error-prone covariates. *American Journal of Epidemiology*, *188*, 222–230. https://doi.org/10.1093/aje/kwy210

Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society*, *76*, 243–263.

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research*, *8*, 2334–2351. https://doi.org/10.1086/sim.3782

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, *29*, 337–346. https://doi.org/10.1002/sim.3782

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

Myers, J. A., Rassen, J. A., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., … Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American Journal of Epidemiology*, *174*(11), 1213–1222. https://doi.org/10.1093/aje/kwr364

Patrick, A. R., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Rothman, K. J., Avorn, J., & Sturmer, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiology and Drug Safety*, *20*(6), 551–559. https://doi.org/10.1002/pds.2098

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/

Ridgeway, G., McCaffrey, D., Morral, A., Griffin, B. A., & Burgette, L. (2017). *Twang: Toolkit for weighting and analysis of nonequivalent groups*. Retrieved from https://CRAN.R-project.org/package=twang

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3), 169-188. https://doi.org/10.1023/A:1020363010465

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20–36. http://dx.doi .org/10.1002/sim.2739

Setoguchi, S., S amd Schneeweiss, Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluation uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology & Drug Safety*, *17*(6), 546–555. https://doi.org/10.1002/pds.1555

Steiner, P. M., Cook, T. D., Shadish, W. R., & Clark, M. H. (2010). The importance of covariate selection in controlling for selectin bias in observational studies. *Psychological Methods*, *15*(3), 250–267. https://doi.org/10.1037/a0018719

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Stuart, E. A, Lee, B. K., & Leacy, F. P. (2013). Prognostic score–based balance measures for propensity score methods in comparative effectiveness research. *Journal of Clinical Epidemiology*, *66*(8), S84–S90. https://doi.org/10.1016/j.jclinepi.2013.01.013

Westreich, D., Lessler, J., & Funk, M. J. (2010). Propensity score estimation: Machine learning and classification methods as alternatives to logistic regression. *Journal of Clinical Epidemiology*, *63*(8), 826–833. https://doi.org/10.1016/j.jclinepi.2009.11.020

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017). *What works clearinghouse: Procedures and standards handbook (version 4.0)*. Retrieved from http://whatworks.ed.gov