Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Yaz* 2018 *Summer* 2018 *Cilt: 9- Sayı: 3* *Volume: 9- Issue: 3*

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: epod@epod-online.org

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

Contact
e-mail: epod@epod-online.org
Web: http://epod-online

Dizinleme / Abstracting & Indexing
DOAJ (Directory of Open Access Journals), TÜBİTAK Ulakbim Sosyal ve Beşeri Bilimler Veri Tabanı, Tei (Türk Eğitim İndeksi)

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Cumhuriyet Üni.)
Akif AVCU (Marmara Üni.)
Asiye Şengül Avşar (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar Şahin Sarkın (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengu BORKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Hacettepe Üni.)
Çiğdem Reyhanlioğlu Keçeoğlu
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Hacettepe Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (ÖSYM)
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.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Halil Özberk (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sutcu Imam Uni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca Usta (Cumhuriyet Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Sakarya Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Sarıçam (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Hacettepe. Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (Hcettepe Üni.)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe 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.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Ragıp Terzi (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Selen Demirtaş ZORBAZ (Ordu Üni)

Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sümeyra SOYSAL (HAcettepe Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of North Carolina)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)

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

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Eylül 2018, Sayı: 9-3)

Journal of Measurement and Evaluation in Education and Psychology (September 2018, Issue: 9-3)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3);239-257

The Interaction Effect of the Correlation between Dimensions and Item Discrimination on Parameter Estimation*

Sakine GÖÇER ŞAHİN** Derya ÇAKICI ESER*** Selahattin GELBAL****

Abstract
There are some studies in the literature that have considered the impact of modeling multidimensional mixed structured tests as unidimensional. These studies have demonstrated that the error associated with the discrimination parameters increases as the correlation between dimensions increases. In this study, the interaction between items' angles on coordinate system and the correlations between dimensions was investigated when estimating multidimensional tests as unidimensional. Data were simulated based on two dimensional, and two-parameter compensatory MIRT model. Angles of items were determined as 0.15º; 0.30º; 0.45º; 0.60º and 0.75º respectively. The correlations between ability parameters were set to 0.15, 0.30, 0.45, 0.60 and 0.75 respectively, which are same with the angles of discrimination parameters. The ability distributions were generated from standard normal, positively and negatively skewed distributions. A total of 75 (5 x 5 x 3) conditions were studied: five different conditions for the correlation between dimensions; five different angles of items and three different ability distributions. For all conditions, the number of items was fixed at 25 and the sample size was fixed at n = 2,000. Item and ability parameter estimation were conducted using BILOG. For each condition, 100 replications were performed. The RMSE statistic was used to evaluate parameter estimation errors, when multidimensional response data were scaled using a unidimensional IRT model. Based on the findings, it can be concluded that the pattern of RMSE values especially for discrimination parameters are different from the existing studies in the literature in which multidimensional tests were estimated as unidimensional.

Key Words: Multidimensional data, unidimensional estimation, correlation, discrimination index.

INTRODUCTION

Unidimensionality, which is one of the most fundamental assumptions of modern measurement theories, refers to measuring a single trait through test. Unidimensionality is necessary for ranking individuals on a scale. On the other hand, unidimensionality assumption is not always met in practice since the measured traits may not be perfectly pure. Thus, the unidimensionality assumption and the item response theory (IRT) models relying on this assumption are criticized in various aspects.

The critics on unidimensionality assumption and structure of tests measuring multiple traits have encouraged researchers to develop and employ multidimensional measurement models. Therefore IRT, which has been used for unidimensional tests from its release until the late 1970s, has been extended to multidimensional tests and has started to be used with the test measuring multiple abilities under the name of multidimensional item response theory (MIRT) since the late 1970s and early 1980s (Ansley & Forsyth, 1985; Reckase, 2009).

Multidimensionality means that the test intends to measure multiple traits. Multidimensionality can be applied with different test structures. In this respect, multidimensional tests may have simple, approximate simple, complex, mixed and semi-mixed structures. A simple structured test consists of multiple subtests each of which measures a single trait, and each item in these subtests is related to a

*An early draft of this paper was presented at International Meeting of Psychometric Society (IMPS) in Beijing, China in 2015.

** Postdoctoral Researcher, University of Wisconsin-Madison, Madison, WI, USA, e-mail: sgocersahin@gmail.com, ORCID ID: orcid.org/0000-0002-6914-354X
***Assit. Prof. Dr., Kırıkkale University, Education Faculty, Kırıkkale, TURKEY, e-mail: deryacakicieser@gmail.com, ORCID ID: orcid.org/0000-0002-4152-6821
****Prof. Dr. Hacettepe University, Education Faculty, Ankara, TURKEY, e-mail: sgelbal@gmail.com, ORCID ID: orcid.org/0000-0001-5181-7262

DOI:10.21031/epod.402992
Received: 07.03.2018
Accepted: 05.07.2018

single trait. Tests with an approximately simple structure are also composed of subtests. Each subtest is approximately unidimensional, which means that there is a dimension that is measured recessively in addition to a dominant dimension (Zhang, 2005; Zhang, 2012). As for the tests with a complex structure, both the entire test and the items in the test are related to more than one ability. From a factor analytic perspective, in complex structured tests, items have factor loadings on multiple abilities (Bulut, 2013; Sheng & Wikle, 2007). Mixed structured tests include both simple and complex items. And the semi-mixed tests include both approximate simple and complex items (Zhang, 2012).

Test dimensionality should be carefully examined before implementation of the tests and analysis and interpretation of results. The implementation and interpretation stages of multidimensional analyses are more complicated than that of unidimensional structures. Stages of multidimensional analyses are more complicated than that of unidimensional structures. Due to convenience of implementing and interpreting the unidimensional IRT models, some researchers lean towards analyses in which multidimensional models are estimated as unidimensional. There are studies in the literature estimating multidimensional tests as unidimensional since 1980s (i.e., Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Harrison, 1986; Kirisci, Hsu, & Yu, 2001, Leucht & Miller; 1992; Reckase, Ackerman, & Carlson, 1988; Zhang, 2008; Zhang, 2012). Estimating multidimensional constructs as unidimensional is generally referred as model misspecification.

There are many studies in the literature about model misspecification. In a study carried out by Drasgow and Parsons (1983), impact of applying unidimensional IRT to multidimensional data on item and person parameters was analyzed using LOGIST program. In the study, conditions, in which medium level heterogenous items were used, fitted better to unidimensional model. In another study carried out by Ansley and Forsyth (1985), parameters acquired from unidimensional estimation of two-dimensional constructs were analyzed. According to the obtained findings, correlations between estimation values and true values of difficulty parameter were higher than the correlation between other parameters. Harrison (1986) analyzed robustness of IRT parameters based on hierarchical factor model under various conditions using LOGIST program. According to these results, it was observed that as the test length increased, estimated and observed values of discrimination index got closer to each other; indicating that LOGIST program created better values for unidimensional constructs; and *D* parameter acquired through this program was more robust to the violation of unidimensionality. With respect to the ability parameter, it has been observed that as the test length increased, and the strength of general factor increased, correlation between ability parameters acquired from unidimensional and multidimensional structures increased and RMSD values decreased. In a study carried out by Reckase, Ackerman, and Carlson (1988), a unidimensional test was attempted to be formed using multidimensional items. Two data sets were used in the study. In the first data set, 80 items were calibrated based on two-parameter logistic model (2 PL). First 20 items of these 80 items were formed to measure only θ_1; second 20 items were formed to measure θ_1 and θ_2 in an equal level; third 20 items were formed to measure only θ_2; and finally, a two-dimensional data set was created as angles of the fourth 20 items could distribute equally between 0 – 90º. According to the simulation results, it was observed that 20 items in the first three groups did not show too much deviation from unidimensionality, and the last 20 items showed better consistence with the multidimensional model. Additionally, it was observed that the whole test showed better fit with the multidimensional model. On the contrary, findings acquired from the real data set showed more different results from the simulation data, and a data set designed as two dimensional with 68 items showed better fit with unidimensional model. In the study carried out by Ackerman (1989), multidimensional data generated based on compensatory and non-compensatory models were calibrated using BILOG and LOGIST programs. According to the results observed using both programs, as the correlation between dimensions in the data generated based on non-compensatory model increased, the correlation of a_1 and a_2 parameters with the estimated a parameter approached to 0. It has been observed that although average absolute errors were a little higher for discrimination and difficulty parameters obtained from BILOG program, errors decreased as the correlation between dimensions increased. It was indicated that *D* parameter was more robust in both programs. Results acquired from non-compensatory model showed similarity with the compensatory model. In addition to this, average absolute errors obtained from BILOG program were lower than the errors obtained from LOGIST program. In a study carried

out by Kirisci, Hsu, and Yu (2001), in cases that unidimensionality and normality assumptions were not met, estimations acquired from BILOG, MULTILOG, and XCALIBRE programs were compared. Test and individual parameters were estimated based on data including three dimensional structures where unidimensional and interdimensional correlation was 0.6 and ability distributions were normal, positively-skewed and platykurtic. RMSE values were used to evaluate the results. RMSE values on the basis of distributions, dimensions, and programs were compared via ANOVA. According to ANOVA results, main effect of distributions and its interaction with other variables were not significant. It was observed that main effect of the dimension was significant only for c_i parameter. In the study where Zhang (2008) analyzed unidimensional parameter estimations and deviations from unidimensionality, used the number of dimensions as four; the test length as 15, 30, and 60; the rate of number of items that load to other dimensions as 20%, 40%, and 60%; and the correlation between factors as 0.00, 0.40, and 0.80. According to the findings, it was observed that as the correlation between secondary dimensions and the dominant dimension increased, the structure did not deviate much from unidimensionality. It was indicated that as the correlation decreased and the rate of items loading to other dimensions increased, the structure diverged from approximate unidimensionality. Another factor affecting divergence from approximate unidimensionality was the test length. When interdimensional correlation was low, shorter tests produced better results compared to longer tests. One of the conditions examined in the studies mentioned above is the structure of the test (approximate simple or complex) while the other most-focused conditions are the skewness of distribution and correlation among the dimensions. In these studies, the general finding about effect of correlation is that when the correlation between dimensions increased, the estimation error was decreased. However, in a study conducted by Gocer Sahin, Walker, and Gelbal (2015), it was reported that contrary to the findings in the literature, especially errors of item parameters increased as the correlation among the dimensions increased and that the lowest level of errors occurred when the correlation was 0.45. In another study carried out by Gocer Sahin (2016), a multidimensional test with a semi-mixed structure was estimated as unidimensional, and the same unexpected pattern related to correlation and test parameters was obtained. A similar study carried out by Kahraman (2013) reported that errors of discrimination increased as the correlation increased when the second dimension of the multidimensional test was ignored and then estimated as unidimensional.

Although there are studies in the literature showed that as the correlation between dimensions increased the estimation errors decreased, in the recent studies an opposite pattern was observed. This may be because of the test structure. In the previous studies, the tests had approximately simple structured items which most of items loaded one factor dominantly and recessively loaded on the second dimension. However, in the recent studies, test structure had mixed format which some items loaded dominantly on one factor some loaded on both dimension. Thus, one factor that makes this study different than others is the test structure. Although the results in the studies conducted by Kahraman (2013), Gocer Sahin, Walker, and Gelbal (2015), Gocer Sahin (2016) appear to be promising, they have not explained the possible reasons behind that results. So, in this study, the focus was on the interaction between correlation and items.

Purpose of the Study

In the recent studies related to the estimations of semi-mixed structured multidimensional tests as unidimensional, we think that increase in errors associated with item parameters because of the increase in correlation between the dimensions may stem from the interaction between the items' angles and the correlation. This study was carried out in order to test whether this hypothesis was true. Therefore, this study aims to answer following questions:

1. How much error is included in parameter estimation when a two-dimensional test is treated as unidimensional?
2. Is there a pattern for error associated with ability parameters in the case of misspecification of two-dimensional tests as unidimensional?

3. How the ability estimations are affected by the interaction among different ability distributions, correlation between dimensions and angles of items on the x-axis?

METHOD

In this study, simulated data sets were used to perform research purpose. Simulation models should be based on realistic situations (Davey, Nering, & Thompson, 1997). In this study the minimum number of items in the large-scale tests was considered test length. In large scale tests for example, in high school entrance exams, each sub test includes 20 questions. So, two dimensional tests with 25 items and with a semi-mixed structure were simulated. According to Hambleton (1989), a large (around 1,000) sample is required to obtain accurate item-parameter estimates in IRT (Hambleton, 1989) for accurate estimates of ability parameter, upon which some high-stakes decisions are made. To eliminate the sample size effect, an enough number of examinees were simulated. In the whole design, the sample size was fixed to be 2,000. The independent variables of the study are correlation among dimensions, items' angle with x-axis, and distribution of ability parameters.

In this respect, the correlation among the ability parameters in the two-dimensional tests is manipulated in an order from the lowest relation to the highest relation (ρ=0.15; ρ=0.30; ρ=0.45; ρ=0.60; ρ=0.75). There are some findings in the literature showing that the shape of distributions affects the parameter estimation in BILOG (Abdel-Fattah, 1994; Kim & Lee, 2014; Kirisci, Hsu, & Yu, 2001; Seong, 1990; Toland, 2008; Yen, 1987). Although it is known that the ability distribution has impacts on the parameter estimation, its impact on semi-mixed structured tests is not known yet. So, in this study ability distribution was one of the independent variables. Since the standard normal distribution is used by default as the initial (prior) ability distribution for calibrating item parameters in BILOG, standard normal distributions were added to the design as a baseline condition. For standard normal distributions, underlying ability distributions for both dimensions were simulated as standard normal $N(0, 1)$. For positive and negative skewed distributions, the values in the Fleisman's (1978) study were used. For positively skewed distributions and negatively skewed distributions skewness and kurtosis were (1.75, 3.75) and (-1.75, 3.75), respectively. For each condition, 100 replications were performed.

In MIRT, items can be represented by item vectors on Cartesian coordinate system. Each item vector is on a line that crosses the origin. The direction of the vector is defined as the vector's angle with positive θ_1 axis. The direction of an i item is calculated through the following equation (Reckase, 2009):

$$\alpha_i = arccos \frac{\alpha_{i1}}{\sqrt{\alpha_{i1}^2+\alpha_{i2}^2}} \quad (1)$$

In Equation 1, a_i refers to the discrimination of item i. Items that are closer to θ_1 axis primarily measure the θ_1 ability while items that are closer to θ_2 axis primarily measure the θ_2 ability. Items have an angle of 45° with both ability axes equally measure both of the abilities (Ackerman, 1994; Ackerman, Gierl, & Walker, 2003). Accordingly, in this study, the angles of item vectors with x axis are manipulated as 15°, 30°, 45°, 60°, and 75°, which are the same numerical values as the correlations. In such a design, the items with angles of 15° and 30° measure the θ_1 ability, the items with angles of 45° measure both θ_1 and θ_2, and the items with angles of 60° and 75° primarily measure the θ_2 ability. Ability parameters were acquired from three different distributions, which were standard normal, positive skewed and negative skewed distribution. In this arrangement, the ability distributions had three conditions, items' angles with x axis had five conditions, and correlations among dimensions had five conditions; which resulted in a total of 75 conditions (3 x 5 x 5). Data were generated through the SAS software on the basis of compensatory two parameter logistic model with the following equation (2) (Reckase, 2009):

$$\mathrm{P}(\mathrm{U_{ij}} = 1 \mid \theta_i, a_i, d_i = \frac{e^{a_i\theta_j^{'}+d_i}}{1+e^{a_i\theta_j^{'}+d_i}} \quad (2)$$

where P is the conditional probability that examinee j's response, U_{ij}, to item i is correct, θ_j is the ability vector, a_i is the discrimination parameter vector, and d_i represents scalar difficulty of item i.

Item and ability parameter estimation were conducted using BILOG.

In order to have a baseline condition for comparison purposes, a unidimensional data set was also simulated. To generate unidimensional data, multidimensional test parameters were utilized. MDISC (maximum discrimination index) and *D* were used as the discrimination and difficulty parameters for unidimensional tests, respectively. *MDISC* is the overall discriminating power of an item which shares the same interpretation as the discrimination parameter in the unidimensional models (Reckase & McKinley, 1991).

$$\mathrm{MDISC}_i = \sqrt{\sum_{\mathrm{k}=1}^{\mathrm{m}} a_{ik}^2} \tag{3}$$

where *m* refers to the number of ability dimensions the a_{ik} variable refers to the discrimination value that belongs to each dimension. The difficulty level of an item is defined as (Reckase, 2009):

$$D_i = \frac{-d_i}{MDISC} \tag{4}$$

In Equation 4, d_i is intercept term. The value of D_i has the same interpretation as the *b* parameter in the unidimensional IRT. The number of items was fixed at 25 and the sample size was fixed at *n* = 2,000 for the simulated unidimensional test data as well. The RMSE values obtained from the unidimensional tests were used as the baseline criterion to evaluate the magnitude of the errors that were obtained from the multidimensional data.

$$RMSE = \sqrt{\frac{\sum_r^n (\hat{X}_{ir} - X_i)^2}{n}} \tag{5}$$

In Equation 5, *i* and *r* represent items (or examinees) and replications, respectively, *n* is the total number of replications, and $\hat{X}_{ir}$ is the estimate of parameter X_i (a_1, a_2 *and* a_{avg} *(the average of* a_1 *and* a_2*)*, *D*, θ_1, θ_2, *and* θ_{avg} *(the average of* θ_1, *and* θ_2*)* or *MDISC*). RMSE (Root Mean Square Error) statistics in the equation (5) were used to evaluate the errors associated with the estimated parameters. This equation is used to calculate the error in ability parameters, and this formula was also adapted to item parameters.

In the findings part, ANOVA was conducted to determine the impact of different correlations, distributions, and angles given in Table 1-7. Although the homogeneity of variances for some data was not met, ANOVA was continued in order to provide consistency in all results. With the aim of comparing the results, Bonferroni's method was used for post hoc comparisons.

RESULTS

a_1 Parameter:

The RMSE values obtained for the a_1 parameter are displayed in Table 1. When the distribution of errors pertaining to the a_1 parameter along the change of the correlations are examined by keeping the item's angle constant, it was observed that the errors decreased as the correlation among the dimensions increased under the conditions with the angles smaller than 45º. Under the conditions where angles were higher than 45º, the errors increased as the correlation among the dimensions increased. The only condition that did not conform to the pattern related to correlation and angle was when the distributions were standard normal, and the angle was 45º.

When the distributions were standard normal, and the item's angle was 45º, then the errors had a hyperbolic curve. In this respect, when the correlation was kept constant, the errors decreased until the angle reached to 45º whereas the errors increased after 45º. An evaluation according to the distributions showed that the skewness of the distributions affected the a_1 parameter. Especially when the items' angles were higher than 45º (when the angles are 60º and 75º), the RMSE values obtained under the conditions of standard normal distributions were higher than the error values obtained under the conditions of skewed distributions. Under other conditions apart from this, the RMSE values obtained

in skewed distributions were bigger than the error values obtained in standard normal distributions. It should also be added that the direction of the skewness had no effect on the a_1 parameter. The important point here is whether the distribution is skewed or standard normal; it is not the direction of the skewness. A comparison of the RMSE values obtained through the estimation of multidimensional data as unidimensional revealed that the errors closest to the criterion values were observed under the conditions where angles were 45º.

a_2 Parameter:

The RMSE values obtained for the a_2 parameter are presented in Table 2. Evaluation of a_2 parameter showed an opposite pattern with a_1 parameter. When the angle was kept constant, errors pertaining to the a_2 parameter increased as the correlation increased in the conditions with the angles smaller than 45º. In the conditions with the angles higher than 45º, the errors decreased as the correlation increased. An evaluation based on the distributions showed that the same symmetric pattern between a_1 and a_2 also occurred. Specifically, when the items' angles were smaller than 45º (when the angles are 15º and 30º), the RMSE values obtained under the conditions of standard normal distribution were higher than the error values obtained under the conditions of other skewed distribution. In the cases that angles were 45º or above, the RMSE values obtained under the conditions of standard normal distribution were lower than the RMSE values obtained under the conditions of skewed distribution. When all these values are compared with the criterion RMSE values, it is observed that in the condition where angle is 45º, the errors related to a_2 parameter were generally lower than the criteria values.

The comparison of the error sizes pertaining to the a_1 and a_2 parameters revealed that in some cases, the errors of a_1 were higher and in other cases, the errors of a_2 were higher. The patterns obtained were generally symmetrical. It is observed that the average error within each condition for both parameters were close to each other.

Table 1. RMSE Values for a_1 Parameter

Angles	Results of Unidimensional data	Correlation of Between Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.058	0.421	0.473	0.472	0.407	0.466	0.465	0.397	0.458	0.456	0.387	0.453	0.444	0.379	0.443	0.435
30^0	0.080	0.307	0.359	0.378	0.273	0.336	0.357	0.238	0.317	0.338	0.211	0.292	0.317	0.184	0.274	0.294
45^0	0.121	0.151	0.245	0.242	0.109	0.219	0.224	0.083	0.206	0.204	0.088	0.193	0.190	0.113	0.183	0.182
60^0	0.101	0.179	0.136	0.151	0.224	0.160	0.173	0.267	0.185	0.196	0.310	0.213	0.225	0.354	0.245	0.251
75^0	0.125	0.533	0.455	0.444	0.564	0.473	0.460	0.595	0.493	0.478	0.626	0.517	0.501	0.655	0.542	0.521

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

Table 2. RMSE Values for a_2 Parameter

Angles	Results of Unidimensional data	Correlation of Between Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.058	0.471	0.434	0.427	0.479	0.436	0.431	0.486	0.441	0.437	0.496	0.442	0.448	0.504	0.449	0.455
30^0	0.080	0.183	0.163	0.173	0.206	0.177	0.181	0.234	0.191	0.192	0.260	0.215	0.207	0.293	0.235	0.227
45^0	0.121	0.146	0.238	0.239	0.107	0.214	0.221	0.080	0.200	0.202	0.086	0.186	0.188	0.115	0.177	0.182
60^0	0.101	0.368	0.457	0.451	0.320	0.434	0.427	0.277	0.410	0.402	0.235	0.384	0.374	0.194	0.357	0.350
75^0	0.125	0.576	0.679	0.692	0.545	0.663	0.678	0.514	0.645	0.660	0.484	0.624	0.639	0.455	0.601	0.620

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

a_{avg} *Parameter*

The RMSE values obtained for the a_{avg} parameter can be seen in Table 3. Under the conditions with standard normal distribution, the highest errors were obtained when the correlation among the dimensions was 0.15, and the lowest errors were obtained when the correlation was 0.45 for the average of *a* parameters. No regular pattern was found under the conditions with standard normal distribution. When the errors are examined for the correlations by keeping the angles fixed, it can be suggested that the errors of a_{avg} yielded a hyperbolic curve for to the correlation between the dimensions. The RMSE values obtained under the conditions with standard normal distribution were generally lower than the values obtained under the conditions with skewed distribution. Under the conditions with skewed distribution, the errors decreased as the correlation among the dimensions increased. When the distributions were skewed, the highest errors were found at 45°, and the lowest errors were found at 15°. The errors closest to the criterion values under the conditions with skewed distribution were obtained when the correlation was 0.75. The sizes of the errors pertaining to the a_{avg} parameter were between the a_1 and a_2 parameters. A comparison of all the obtained values with the criterion RMSE values showed that the errors, which were obtained when the correlation among the dimensions was 0.45 and the distribution was standard normal, were generally lower than the criterion values.

MDISC Parameter:

The RMSE values obtained for the *MDISC* parameter are presented in Table 4. It is observed that the *MDISC* parameter which corresponds to the discrimination parameter in the unidimensional IRT included more errors than all other discrimination parameters. The error values decreased as the correlation increased. In general, the errors increased as the angles increased. Under each condition of distribution, the lowest errors were obtained when the correlation was 0.75. The RMSE values obtained under the conditions of standard normal distribution were lower than the error values obtained under the conditions of skewed distribution. Whether the distribution is right or left skewed is not very influential on the RMSE. Accordingly, the effective condition for the RMSE is whether the distribution is standard normal or not. In general, it can be suggested that, the errors pertaining to the *MDISC* were quite higher than the criterion values.

Table 3. RMSE Values for a_{avg} Parameter

Angles	Results of Unidimensional data	Correlation of Between Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.058	0.081	0.116	0.098	0.070	0.105	0.089	0.066	0.096	0.080	0.068	0.086	0.077	0.076	0.080	0.074
30^0	0.080	0.103	0.156	0.183	0.072	0.138	0.164	0.051	0.125	0.149	0.055	0.112	0.137	0.082	0.109	0.126
45^0	0.121	0.139	0.236	0.234	0.094	0.210	0.216	0.062	0.196	0.196	0.069	0.182	0.182	0.101	0.172	0.174
60^0	0.101	0.113	0.207	0.205	0.073	0.190	0.188	0.056	0.174	0.170	0.068	0.160	0.156	0.102	0.151	0.148
75^0	0.125	0.061	0.173	0.184	0.058	0.164	0.176	0.071	0.158	0.167	0.092	0.154	0.159	0.116	0.152	0.155

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

Table 4. RMSE Values for *MDISC* Parameter

Angles	Results of Unidimensional data	Correlation of Between Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.058	0.463	0.515	0.513	0.448	0.508	0.506	0.438	0.500	0.497	0.427	0.494	0.484	0.419	0.484	0.475
30^0	0.080	0.466	0.514	0.539	0.427	0.489	0.516	0.387	0.467	0.495	0.354	0.437	0.471	0.315	0.413	0.444
45^0	0.121	0.551	0.636	0.636	0.449	0.601	0.611	0.443	0.576	0.580	0.391	0.547	0.551	0.345	0.510	0.523
60^0	0.101	0.551	0.636	0.633	0.501	0.611	0.609	0.457	0.585	0.582	0.414	0.557	0.552	0.370	0.526	0.527
75^0	0.125	0.627	0.729	0.743	0.596	0.713	0.728	0.565	0.695	0.711	0.535	0.673	0.689	0.505	0.650	0.670

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

D parameter:

The RMSE values obtained for the *D* parameter are displayed presented in Table 5. As for the errors pertaining to the difficulty parameter obtained when the two-dimensional tests were estimated as unidimensional, it was observed that the errors increased as the correlation among the dimensions increased. In the case of standard normal distributions, the lowest error occurred when the correlation among the dimensions was 0.15 while the highest error occurred when the correlation was 0.75. However, no regular pattern was found regarding the errors under the condition with skewed distributions. Accordingly, in the case that distributions were skewed, and the angle was 15º and 75º, the errors decreased as the correlation increased. When the item's angle with the *x* axis was 30º, 45º and 60º, and the distribution was positively-skewed, RMSE values again produced a hyperbolic curve. Accordingly, errors decreased until the correlation of 0.45 and they increased again after the correlation of 0.45. The pattern that was obtained in the positively-skewed distribution was generally observed in the negatively-skewed distribution. When the correlations and distributions were fixed, and the angles increased, the errors did not exhibit a regular pattern. Under the condition with correlation of 0.15 between the dimensions and when the distribution was standard normal, considering the errors pertaining to the *b* parameter showed that the criterion values were closest to each other. Under this condition, almost all of the errors that were obtained by estimating the two-dimensional structures as unidimensional were lower than the criterion value.

θ_1 parameter:

The RMSE values obtained for the θ_1 parameter are presented in Table 6. Errors pertaining to the θ_1 parameter were affected by both correlation between ability parameters and angle of items. In this respect, the errors decreased as the correlation between the dimensions increased. In the case that distributions and correlations were held constant, the errors increased only when the angles increased. Specifically, the increase of the angle under the conditions of low correlation resulted in a significant increase in the errors; the increase of the angle under the conditions of high correlation had relatively lower effect on the errors. The highest errors were obtained when the correlation was 0.15 and the angle was 75º. Varying the distribution did not have a significant effect on the errors. Under all conditions, the errors obtained in standard normal distribution had lower values than in the positively and negatively skewed distributions. The errors acquired from the skewed distributions under the same conditions had similar values. The errors obtained for the θ_1 parameter were quite higher than the criterion values under all conditions. When the correlation was 0.75, the criterion RMSE and the obtained RMSE values were closest to each other, but the difference increased as the angle increased.

Table 5. RMSE Values for *D* Parameter

Angles	Results of Unidimensional data	Correlation of Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.053	0.095	0.180	0.182	0.100	0.171	0.187	0.120	0.164	0.193	0.139	0.158	0.200	0.160	0.157	0.214
30^0	0.081	0.078	0.179	0.213	0.108	0.169	0.191	0.151	0.164	0.181	0.191	0.171	0.176	0.230	0.182	0.179
45^0	0.123	0.078	0.208	0.209	0.124	0.190	0.196	0.175	0.189	0.193	0.222	0.193	0.196	0.261	0.204	0.200
60^0	0.090	0.076	0.200	0.197	0.109	0.187	0.178	0.151	0.178	0.164	0.189	0.179	0.165	0.225	0.187	0.170
75^0	0.095	0.057	0.222	0.247	0.068	0.201	0.229	0.087	0.193	0.217	0.108	0.180	0.199	0.131	0.170	0.183

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

Table 6. RMSE Values for θ_1 Parameter

Angles	Results of Unidimensional data	Correlation of Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.053	0.447	0.493	0.494	0.439	0.489	0.489	0.431	0.482	0.484	0.421	0.472	0.474	0.410	0.467	0.464
30^0	0.081	0.597	0.641	0.633	0.560	0.612	0.607	0.519	0.579	0.576	0.477	0.541	0.542	0.432	0.497	0.500
45^0	0.123	0.748	0.776	0.777	0.685	0.731	0.733	0.618	0.679	0.682	0.548	0.618	0.619	0.472	0.549	0.548
60^0	0.090	0.930	0.945	0.951	0.842	0.881	0.888	0.753	0.753	0.813	0.656	0.725	0.731	0.551	0.626	0.627
75^0	0.095	1.108	1.122	1.121	1.006	1.044	1.045	0.895	0.954	0.958	0.776	0.845	0.850	0.638	0.717	0.719

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

θ_2 parameter:

The RMSE values obtained for the θ_2 parameter are presented in Table 7. As seen in Table 7, errors pertaining to the θ_2 parameter significantly decreased as the correlation between dimensions increased. It can be suggested that the varying the distribution did not affect the errors significantly. When the distributions are compared to each other with other conditions being fixed, the lowest error values were obtained under the condition of standard normal distribution. Errors obtained in positively and negatively-skewed distributions under the same conditions were close to each other in general. As the angles increased, the errors obtained for θ_2 decreased. When all the results are considered together, it was observed that the lowest error occurred when the correlation was 0.75 and the angle was 75º, and the highest error occurred when the correlation was 0.15 and the angle was 15º. The difference between the criterion values and the estimated values for the θ_2 parameter increased as the angles and correlations increased; under all conditions, the criterion RMSE values were lower than the RMSE values obtained for the multidimensional data.

When the two-dimensional structures are estimated as unidimensional, the errors pertaining to the θ_2 parameter had similarities to the error values obtained for θ_1 under the same conditions. According to this, the errors were affected by the increase of the correlation and by the distributions in the same way. However, contrary to the situation observed in the θ_1 parameter, the errors of θ_2 decreased as the angle increased. The error patterns obtained for θ_1 and the error patterns obtained for θ_2 were opposite. In this respect, it can be suggested that the errors obtained for θ_1 and θ_2 when the total of the angles were 90º were very close to each other. The error of θ_1 under the condition of 15º angle was very close to the error of θ_2 under the condition of 75º. Similarly, the error of θ_1 under the condition of 30º angle was very close to the error of θ_2 under the condition of 60º angle. Therefore, the errors obtained for both θ_1 and θ_2 under similar conditions and under the condition of 45º angle were close to each other.

θ_{avg} parameter:

The RMSE values obtained for the θ_{avg} parameter are presented in Table 8. Table 8 demonstrates the errors pertaining to the θ_{avg} parameter, which is the average of the θ_1 and θ_2 parameters. According to the table, the variations in angles and correlations affected the errors pertaining to the θ_{avg} parameter. However, this effect was not as high as in θ_1 and θ_2; yet, it was lower. Similarly, the errors decreased as the correlation increased. The increase of the angles had a varying effect on the errors. Accordingly, under all conditions, the errors initially decreased and then increased as the angles increased. The lowest errors were obtained under the conditions of 45º angles. Variation in distributions did not significantly affect the error of θ_{avg}. Errors obtained in standard normal distribution had the lowest values while similar errors were obtained in positively and negatively-skewed distributions. This finding is similar to the one found for θ_1 and θ_2. The criterion RMSE values were found to be lower than the RMSE values obtained for multidimensional tests under all conditions. The condition in which the criterion values and the errors pertaining to the multidimensional data was closest to each other when the angles were 45º.

ANOVA results about the comparison of results

According to ANOVA results, the average errors of discrimination parameter varied in accordance with distributions (for a_1 [$F_{2,7497}$=16.700, p<.05]; for a_2 [$F_{2,7497}$=150.015, p<.05]; for a_{avg} [$F_{2,7497}$=2960.506, p<.05]; for MDISC [$F_{2,7497}$=1679.966, p<.05]). Based on the results of post hoc comparisons, there was not any significant difference between errors obtained under positively and negatively skewed distribution conditions for a_1 and a_2, and the errors obtained under normal conditions were smaller. For MDISC and a_{avg}, errors obtained for all distribution conditions were different from each other; the lowest error values were obtained under standard normal distribution and the highest error values were obtained under negatively skewed distribution.

According to ANOVA results, the average errors of discrimination parameter varied by interdimensional correlation (for a_1 [$F_{4,7495}$=3.754, p<.05]; for a_2 [$F_{4,7495}$=3.279, p>.05]; for a_{avg} [$F_{4,7495}$=149.596, p<.05]; for *MDISC* [$F_{4,7495}$=224.635, p<.05]). Based on the conducted post hoc comparisons, for a_1, there was a significant difference only between errors obtained in correlation of 0.15 and 0.75. According to this, error values obtained under 0.15 correlation condition were lower. For a_2, it was observed that the errors obtained under the condition where correlation was 0.30 were higher than the errors obtained under the conditions where correlations were 0.15 and 0.75. No significant difference was obtained among the errors apart from other conditions. For a_{avg} and *MDISC*, errors obtained under all correlation conditions were not different from each other. According to this, the highest errors were obtained in 0.15 correlation value, and the lowest errors were obtained in 0.75 correlation value.

It was determined that the average errors of discrimination parameter varied by angles (for a_1 [$F_{4,7495}$=9211.581, p<.05]; for a_2 [$F_{4,7495}$=7896.183, p<.05]; for a_{avg} [$F_{4,7495}$=736.080, p<.05]; for *MDISC* [$F_{4,7495}$=1372.812, p<.05]). Based on the results of post hoc test, errors obtained from all angles were different from each other. When means were examined, for a_1 and a_2, errors got lower up to 45º, had the lowest value at 45º, and got higher after 45º. For MDISC, as angles increased errors also increased; and for a_{avg}, a systematic pattern couldn't be obtained.

According to the results of ANOVA carried out for *D* parameter, the average errors of this parameter varied by distributions [$F_{2,7497}$=917.760, p<.05]. Based on the results of post hoc test, errors obtained from all correlations were different from each other. When means were examined, it was observed that errors obtained under negatively skewed distribution conditions were the highest, and errors obtained under standard normal distribution conditions were the lowest.

According to the results of ANOVA conducted for *D* parameter, the average errors of this parameter varied by interdimensional correlation [$F_{4,7497}$=81.988, p<.05]. Base on the results of post hoc comparisons, errors obtained from all correlation values were different from each other. When means were examined, in general, as interdimensional correlation increased, errors also increased.

Finally, it was determined that the average errors of *D* parameter varied by angles [$F_{4,7495}$=69.682, p<.05]. Based on the results of post hoc test, only the errors under conditions in which the angles were 30º and 60º were not different from each other. Errors obtained under all other conditions were different from each other.

According to the results of ANOVA, it was determined that errors of ability parameter varied by distributions (for θ_1 [$F_{2,7497}$=67.582, p<.05]; for θ_2 [$F_{2,7497}$=61.608, p<.05]; for θ_{avg} [$F_{2,7497}$=344.435, p<.05]). Based on the results of post hoc comparisons, for ability parameter, there was not any difference in positively and negatively skewed distributions; errors obtained under standard normal distribution conditions were lower.

According to the results of ANOVA, the errors of ability parameter varied by correlations (for θ_1 [$F_{4,7495}$=448.577, p<.05]; for θ_2 [$F_{4,7495}$=349.489, p<.05]; for θ_{avg} [$F_{4,7495}$=310.452, p<.05]). Based on the results of post hoc comparisons, errors obtained from all correlation values were different from each other. When means were analyzed, as interdimensional correlation for all ability parameters under all conditions increased, errors decreased.

Finally, according to the results of ANOVA, the average errors of ability parameter varied by angles (for θ_1 [$F_{4,7495}$=4737.972, p<.05]; for θ_2 ([$F_{4,7495}$=6193.641, p<.05]; for θ_{avg} [$F_{4,7495}$=4705.022, p<.05]). Based on the results of post hoc comparisons, errors obtained from all correlation values were different from each other. When means were analyzed, it was observed that for θ_1, as angles increased, errors also increased; for θ_2 and θ_{avg}, as angles increased, errors decreased.

Table 7. RMSE Values for θ_2 Parameter

Angles	Results of Unidimensional data	Correlation of Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.053	1.157	1.173	1.171	1.055	1.099	1.088	0.940	1.007	0.998	0.820	0.902	0.891	0.688	0.770	0.759
30^0	0.081	0.927	0.935	0.945	0.842	0.876	0.884	0.753	0.804	0.818	0.660	0.721	0.734	0.557	0.625	0.636
45^0	0.123	0.751	0.779	0.779	0.687	0.732	0.734	0.621	0.677	0.680	0.551	0.620	0.622	0.474	0.550	0.551
60^0	0.090	0.574	0.621	0.616	0.537	0.595	0.591	0.498	0.498	0.560	0.456	0.525	0.525	0.410	0.483	0.481
75^0	0.095	0.414	0.475	0.477	0.402	0.467	0.470	0.387	0.454	0.458	0.372	0.441	0.445	0.355	0.428	0.431

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

Table 8. RMSE Values for θ_{avg} Parameter

Angles	Results of Unidimensional data	Correlation of Abilities														
		ρ_1=0.15			ρ_1=0.30			ρ_1=0.45			ρ_1=0.60			ρ_1=0.75		
		SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***	SND*	PSD**	NSD***
15^0	0.053	0.586	0.604	0.601	0.550	0.581	0.573	0.511	0.555	0.546	0.475	0.527	0.519	0.442	0.499	0.489
30^0	0.081	0.426	0.447	0.453	0.401	0.431	0.439	0.379	0.419	0.428	0.364	0.409	0.419	0.351	0.404	0.414
45^0	0.123	0.367	0.400	0.399	0.347	0.387	0.390	0.330	0.383	0.383	0.320	0.377	0.381	0.314	0.380	0.381
60^0	0.090	0.414	0.439	0.442	0.386	0.424	0.427	0.363	0.363	0.414	0.345	0.402	0.404	0.333	0.396	0.396
75^0	0.095	0.527	0.545	0.546	0.486	0.519	0.521	0.448	0.494	0.496	0.411	0.465	0.469	0.376	0.438	0.442

*SND: Standard Normal Distribution, **PSD: Positive Skewed Distribution, ***NSD: Negative Skewed Distribution

DISCUSSION and CONCLUSION

The studies in the literature have suggested that errors pertaining to discrimination parameter increase as the correlation between the dimensions increases (Ansley & Forsyth, 1985; Ackerman, 1989; Zhang, 2008). In this study, *MDISC*, one of the discrimination parameters, displayed such a pattern. In addition to *MDISC*, the errors pertaining to the a_1 parameter under the conditions that items' angles were smaller than 45º were in line with these studies in the literature, and an opposite pattern was observed on the error values for the conditions with angles, higher than 45º. Since the a_2 parameter had an opposite pattern with a_1, the a_2 parameter under the conditions of angles larger than 45º is in line with these studies in the literature, and the errors decreased as the correlation among the dimensions increased under these conditions. Thus, it can be suggested that in this study, the most noticeable value especially for the a_1 and a_2 parameters was the 45 point (45º angle and 0.45 correlation). The RMSE values calculated for a_{avg}, which is the average of the a_1 and a_2 parameters, showed a different pattern than the existing studies' values in the literature. Accordingly, the lowest errors for a_1, a_2 and a_{avg} were generally obtained under the conditions in which the angle was 45º, and the errors pertaining to the a_1 and a_2 parameters produced a hyperbolic curve when the correlations were kept constant. Gocer Sahin, Walker, and Gelbal (2015) and Gocer Sahin (2016) reported that the average angles of the items they used for their studies were around 45º. The errors pertaining to the discrimination parameter produced a hyperbolic curve in these authors' studies, too. In this respect, the findings obtained in this study are in line with the studies of Gocer Sahin, Walker, and Gelbal (2015) and Gocer Sahin (2016). If the angles were bigger than 45º, then the errors increased as the correlation increased. And, this finding was consistent with the findings of Kahraman (2013). All the discussions above are valid for the conditions in which distributions are standard normal; while the pattern obtained in skewed distributions is similar to the one in the standard normal distribution, the conditions in which the lowest RMSE values were obtained in skewed distributions are different.

Although the pattern of the a_1 and a_2 parameters were found to be contrary to previous studies in the literature, the *MDISC* parameter had a pattern that is similar to the ones reported in the studies of Ansley and Forsyth (1985), Ackerman (1989), Zhang (2008), Gocer Sahin, Walker, and Gelbal (2015), Gocer Sahin (2016). According to findings, the errors decreased as the correlation among the dimensions increased. Besides, as the angles of the items increased, (i.e. as the complexity of the items increased), the RMSE values increased. This is an expected result since *MDISC* corresponds to the discrimination of the multidimensional IRT model when it is considered a unidimensional IRT model.

With the 45º angle being the breakpoint, when the angles for a_1 were higher than 45º (when the angles are 60º and 75º), the RMSE values obtained under the conditions of standard normal distribution were found to be higher than the errors obtained under the conditions of skewed distribution. When the angle for a_1 was 45º or smaller, the error values obtained in conditions with the skewed ability distributions were higher. The pattern for the a_2 parameter was exactly the opposite of this pattern. It can be suggested that the a_1 and a_2 parameters were not generally affected by the skewed distribution. Although skewed distributions did not affect a_1 and a_2 parameters, the a_{avg} and *MDISC* parameters were affected by skewed distributions. The RMSE values obtained for the a_{avg} and *MDISC* parameters under all conditions of standard normal distribution were lower than the RMSE values obtained under the conditions of skewed distribution, but this difference was not very large. It was also mentioned in the study of Kirisci, Hsu, and Yu (2001) that especially the *MDISC* parameter was not affected by skewed distributions. In the studies of Gocer Sahin, Walker, and Gelbal (2015) and Gocer Sahin (2016), in which the distributions were manipulated as standard normal or only normal, it was reported that the mentioned distributions did not affect the discrimination parameter.

45º angle and 0.45 correlations can be suggested to be the critical values for the discrimination parameters of the tests with a semi-mixed structure, especially for the a_1, a_2 and a_{avg} parameters. If a test parameter with few errors is desired in the estimation of a multidimensional test with a semi-mixed structure as unidimensional, then it can be recommended to use a test in which the items' angles are 45º. If the correlation is 0.45 in such a test, then it is possible to obtain minimum errors.

As for the errors pertaining to the difficulty parameter obtained when the two-dimensional tests were estimated as unidimensional, it is observed that the errors increased as the correlation among the dimensions increased. In the case of standard normal distributions, the lowest error occurred when the correlation among the dimensions was 0.15 while the highest error occurred when the correlation was 0.75. However, no regular pattern was found regarding the errors under the condition of skewed distributions. Almost all of the errors that were obtained by estimating the two-dimensional structures as unidimensional were lower than the criterion value.

The errors obtained for difficulty parameter were generally lower than errors of other parameters. According to that result it can be concluded that difficulty parameter is the robust parameter. This result is similar to the literature. It did not matter whether the distribution was positively or negatively skewed for the difficulty parameter; instead, the main concern was whether the distribution was standard normal or not.

The errors for ability parameters increased as correlation between dimensions increased. This result is similar to the literature (Ackerman, 1989; Ansley & Forsyth, 1985; Doody, 1985; Drasgow & Parsons, 1983; Gocer Sahin, 2016; Zhang, 2008). Interestingly, although items' angles increased the RMSE decreased for θ_1, and although items' angles decreased the RMSE increased for θ_2. It did not matter whether the distribution was positively or negatively skewed for the ability parameters; instead, the main consideration was whether the distribution was standard normal or not. Because when the distributions were skewed, higher errors were obtained than standard normal distributions. This result is similar to the literature. For example, in Gocer Sahin's (2016) study, errors for θ_{avg} were between the errors for θ_1 and θ_2.

Limitations and Suggestions

This study is limited by its research design that has two dimensional data, and two-parameter logistic and compensatory model. The generalizability of the results is limited to the studied conditions; which were a test with 25 items, a sample size with 2,000 examinees, correlations between dimensions with 0.15, 0.30, 0.45, 0.60 and 0.75; angles that items have with *x* axis are 15^o, 30^o, 45^o, 60^o and 75^o; and lastly, distributions which were standard normal, positively skewed and negatively skewed. Another limitation of this study is that the results are based on only one software. Multiple software programs may result in differences in parameter estimates. In this study only RMSE statistics was used to evaluate the results. Bias or other statistics could also be calculated for this purpose.

Based on the conditions of this study, a multidimensional test which has a high correlation between dimensions is suggested for the researchers who aim to scale the abilities of individuals to a one-level scale. However, if the aim is to develop a qualified test, for a two-dimensional test, items that have 0.45 interdimensional correlation and have 45^o angles with *x* axis should be used. If the estimation is carried out through BILOG program, ability distribution should be standard normal or normal.

REFERENCES

Abdel-fattah, A. A. (1994, April). *Comparing BILOG and LOGIST estimates for normal, truncated normal, and beta ability distributions.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*(1), 37-48. https://doi.org/10.1177/014662168500900104

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-27. https://doi.org/10.1177/014662168901300201

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

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Doctoral Dissertation). Available from ProQuest Dissertations and Theses database (UMI No. 3589000).

Davey, T., Nering M. L. & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report 97-4). Iowa City, IA: ACT, Inc.

Doody, E. N. (1985, April). *Examining the effects of multidimensional data on ability and item parameter estimation using the three-parameter logistic model*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7*, 189-199. https://doi.org/10.1177/014662168300700207

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532. https://doi.org/10.1007/BF02293811

Gocer Sahin, S. (2015). *Yarı karışık yapılı çok boyutlu yapıların tek boyutlu olarak ele alınması durumunda kestirilen parametrelerin incelenmesi* (Doctoral Dissertation). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Gocer Sahin S., Walker, C. M., & Gelbal, S. (2015). The Impact of model misspecification with multidimensional test data. In L. A. van der Ark, D. M. Bolt, S. M. Chow, J. A. Douglas & W. C. Wang (Eds.), *Quantitative Psychology Research: The 79th Annual Meeting of the Psychometric Society* (pp.133-44). New York, NY: Springer.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115. doi: 10.2307/1164972

Kahraman, N. (2013) Unidimensional interpretations for multidimensional test items. *Journal of Educational Measurement, 50*(2), 227-246. https://doi.org/10.1111/jedm.12012

Kim, K. Y., & Lee, W. (2014, July). *Recovery of item parameters under various IRT item calibration methods*. Paper presented at the International Meeting of the Psychometric Society, Madison, WI.

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*(2), 146-162. https://doi.org/10.1177/01466210122031975

Leucht, R. M & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16*, 279-293. https://doi.org/10.1177/014662169201600308

Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences)*. New York: Springer.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193-203. https://doi.org/10.1111/j.1745-3984.1988.tb00302.x

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*(4), 361-373. https://doi.org/10.1177/014662169101500407

Seong, T-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement, 14*(3), 299-311. https://doi.org/10.1177/014662169001400307

Sheng Y. & Wikle C. K. (2007). Comparing multidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 68*(3), 413-430. https://doi.org/10.1177/0013164406296977

Toland, M. (2008). *Determining the accuracy of item parameter standard error of estimates in BILOG-MG 3*. (Doctoral Dissertation). Avaible from ProQuest Dissertations and Theses database (UMI No. 3317288).

Yen, M. Y. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimension. *The Journal of Experimental Education, 77*(2), 147-166.

Zhang, J. (2005). *Estimating multidimensional item response models with mixed structure* (ETS Research Report 05-04). Princeton, NJ: Educational Testing Service.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*(5), 375-398. https://doi.org/10.1177/0146621612445904

Boyutlar Arası Korelasyon ile Madde Ayırt Ediciliği Arasındaki Etkileşimin Parametre Kestirimi Üzerine Etkisi

Giriş

Testlerin uygulanması, verilerin analizi ve yorumlanmasından önce test boyutluluğunun titizlikle incelenmesi gerekir. Tek boyutluluk sayıltısının MTK için bu denli önemli olması ve tek boyutluluğa dayanan modellerin uygulanması ve yorumlanmasının daha kolay olması, araştırmacıları çok boyutlu modellerin tek boyutlu olarak ele alındığı çalışmalara yöneltmektedir. Çok boyutlu testlerin tek boyutlu olarak kestirilmesi ile ilgili çalışmaların 1980'li yıllardan itibaren yapıldığı görülmektedir. Bu tür çalışmalar genel olarak modelin yanlış tanımlanması (model misspecification) olarak adlandırılmaktadır.

Modeli yanlış tanımlama çalışmalarında incelenen koşullardan biri testin yapısı olup (yaklaşık basit veya karmaşık) bunun dışında en çok ele alınan koşullar, boyutlar arası korelasyon ve dağılımların çarpıklığıdır (Ackerman, 1989; Ansley ve Forsyth; 1985; Drasgow ve Parsons; 1983; Harrison, 1986; Kirisci, Hsu ve Yu, 2001, Leucht ve Miller, 1992; Reckase, Ackerman ve Carlson, 1988; Zhang, 2008; Zhang, 2012). Kahraman (2013) tarafından yapılan bir çalışmada, çok boyutlu bir testin ikinci boyutunun ihmal edilerek tek boyutlu kestiriminde, korelasyon arttıkça ayırt ediciliğe ait hatanın arttığı belirtilmiştir.

Son yıllarda yapılan çalışmalarda yarı karışık (semi-mixed) yapılı çok boyutlu testlerin tek boyutlu olarak kestirilmesinde, boyutlar arası korelasyon arttıkça madde parametrelerine ait hataların da artmasının, maddelerin analitik düzlemdeki açıları ile boyutlar arasındaki korelasyonun etkileşiminin bir sonucu olduğu düşünülmektedir. Bu çalışma, bu hipotezin doğru olup olmadığını test etmek üzere yapılmıştır. Dolayısıyla bu çalışmanın amacı, iki boyutlu testlerin tek boyutlu olarak ele alınması durumunda kestirilen parametrelerin, farklı yetenek dağılımları, boyutlar arası korelasyon ve maddenin x ekseni ile yaptığı açı değişkenlerinin kombinasyonlarından nasıl etkilendiğini belirlemektir.

Yöntem

Bu çalışmada, bir testte yer alan maddelerin x ekseniyle yaptığı açılar ile boyutlar arası korelasyonlar manipüle edilerek, boyutlar arası korelasyon ile maddelere ait açıların etkileşiminin parametre kestirimi üzerine etkisi incelenmiştir. Çalışmada simülasyon yoluyla yarı karışık yapılı, 25 maddeden oluşan iki boyutlu testler üretilmiştir. Tüm desende örneklem büyüklüğü 2000 olacak şekilde sabitlenmiştir. Ele alınan iki boyutlu testlerde yetenek parametreleri arasındaki korelasyon düşük ilişkiden yüksek ilişkiye doğru sıralanacak biçimde (ρ=0,15; ρ=0,30; ρ=0,45; ρ=0,60; ρ=0,75) değişimlenmiştir.

Bu çalışmada madde vektörlerinin x ekseniyle yaptığı açı, korelasyonlar ile aynı sayısal değerlerde olmak üzere 15^o, 30^o, 45^o, 60^o ve 75^o şeklinde manipüle edilmiştir. Bu şekilde oluşturulan desende açıları (15^o ve 30^o) olan maddeler öncelikli olarak θ_1 yeteneğini, açıları 45^o olan maddeler hem θ_1 hem θ_2 yeteneğini ve açıları 60^o ve 75^o olan maddeler ise öncelikli olarak θ_2 yeteneğini ölçmektedir. Yetenek parametreleri ise standart normal, sağa çarpık ve sola çarpık dağılım olmak üzere üç farklı dağılımdan elde edilmiştir.

Bu şekilde düzenlenen çalışmada yetenek dağılımları 3; maddelerin x ekseniyle yaptığı açılar 5 ve boyutlar arası korelasyon 5 koşul olmak üzere toplam (3 x 5 x 5) 75 hücreli bir desen oluşturulmuştur. Veriler, SAS programı aracılığıyla telafisel, 2 parametreli lojistik modele dayanarak üretilmiştir. Veri üretiminde 100 replikasyon yapılmıştır.

Çok boyutlu yapıların tek boyutlu olarak ele alınması durumunda kestirilen parametrelerin içerdiği hataların değerlendirilmesinde RMSE istatistiğinden faydalanılmıştır. RMSE değerleri, tüm parametreler için ayrı ayrı hesaplanmıştır.

Çalışmada çok boyutlu testler dışında gerçekte tek boyutlu olan 25 maddeli ve 2000 kişilik bir test tek boyutlu olarak kestirilmiştir. Tek boyutlu test oluştururken, çok boyutlu testlere ait parametrelerden yararlanılmıştır. Buna göre çok boyutlu testlere ait *MDISC* ve *D* parametresi, tek boyutlu teste ait gerçek *a* ve *b* parametrelerini oluşturmuştur.

Sonuç ve Tartışma

Literatürde yapılan çalışmalarda boyutlar arası korelasyon arttıkça ayırt ediciliğe ait hataların azaldığı belirtilmiştir (Ackerman,1989; Ansley ve Forsyth, 1985; Zhang, 2008). Bu çalışmada ise ayırt edicilik parametrelerinden *MDISC*'in bu örüntüye sahip olduğu görülmüştür. *MDISC*'in yanı sıra maddelerin açılarının 45º'den küçük olduğu koşullarda a_1 parametresine ait hatalar alan yazındaki bu çalışmalar ile paralellik göstermekte, boyutlar arası korelasyon arttıkça hatalar azalmaktadır. a_2 parametresi a_1 ile ters bir örüntü göstermiştir. Bu çalışmada özellikle a_1 ve a_2 parametreleri için en önemli değerin 45 noktası (45º'lik açı ve 0,45 korelasyon) olduğu söylenebilir. a_1 ve a_2 bu iki parametrenin ortalaması olan a_{ort} için hesaplanan RMSE değerleri alan yazından farklı bir örüntü göstermiştir. Çarpık dağılımlarda elde edilen örüntü standart normal dağılım ile benzer olmakla birlikte çarpık dağılımlarda en düşük RMSE değerlerinin elde edildiği koşullar farklılık göstermektedir.

Bu çalışmada a_1 için açıların 45º'den (açılar, 60º ve 75º) yüksek ve dağılımın standart normal olduğu koşullarda elde edilen RMSE değerleri, çarpık dağılım koşullarında elde edilen hatalardan daha yüksek olmakla beraber bu fark çok fazla değildir. a_1 için açı 45º ve 45º'den küçükken çarpık dağılımlarda elde edilen hata değerleri daha yüksektir. Bu durum a_2 parametresi için tam tersidir. Ancak yine de genel olarak çarpık dağılımın a_1 ve a_2 parametresini etkilemediği söylenebilir. Her ne kadar çarpık dağılımlar a_1 ve a_2 parametrelerini etkilemese de a_{ort} ve *MDISC* parametreleri çarpık dağılımlardan etkilenmektedir. Dağılımın standart normal olduğu bütün koşullarda a_{ort} ve *MDISC* parametreleri için elde edilen RMSE değerleri çarpık dağılım koşullarındaki RMSE değerlerinden düşüktür.

Yarı karışık yapılı testler için özellikle a_1, a_2 ve a_{ort} parametrelerine ilişkin açının 45º ve boyutlar arası korelasyonun 0,45 olduğu koşulların kritik RMSE değerine sahip olduğu söylenebilir. Buna göre çok boyutlu yarı karışık yapılı bir test tek boyutlu olarak kestirildiğinde, madde açılarının 45º olduğu testlerde test parametresinin düşük miktarda hata içerdiği görülmüştür. Bu test ile beraber boyutlar arası korelasyon 0,45 olduğunda ise hatalar en düşük değerlerini almıştır.

Güçlük parametresi için elde edilen hata değerleri, diğer parametrelerinkinden genel olarak daha azdır. Buna göre bu çalışmada da alan yazına benzer olarak güçlük parametresinin daha dayanıklı olduğu söylenebilir. Güçlük parametresi için de dağılımın sağa veya sola çarpık olması önemli olmayıp; dağılımın standart normal olması veya olmaması önemlidir.

Yetenek parametrelerine ait hatalar, boyutlar arası korelasyon arttıkça azalmıştır. Bu bulgu alan yazındaki benzer çalışmalar ile paraleldir (Ackerman, 1989; Ansley & Forsyth, 1985; Doody, 1985; Drasgow & Parsons, 1983; Gocer Sahin, 2016; Zhang, 2008). θ_1 için maddelerin açıları arttıkça hatalar artmasına rağmen, θ_2 için açı arttıkça hatanın azalması ilginç bir sonuçtur. Yetenek parametreleri için dağılımın sağa veya sola çarpık olması önemli olmamakla birlikte dağılımın standart normal olması önemli bir koşuldur. Çünkü dağılım çarpıklaştığında yetenek parametrelerine ait hatalar artmaktadır. Bu durum alan yazın ile benzerlik göstermektedir. Gocer Sahin (2016)'nın çalışmasına benzer olarak θ_{ort} için elde edilen hata değerleri θ_1 ve θ_2 için elde edilen hataların arasında değer almıştır.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3);258-276

An Implementation of the Gibbs Sampling Method under the Rasch Model

Sedat ŞEN* Tuğba KARADAVUT** Hyo Jin EOM***
Allan S. COHEN**** Seock-Ho KIM*****

Abstract
A brief explication of the implementation of the Gibbs sampling method via rejection sampling to obtain Bayesian estimates of difficulty and ability parameters under the Rasch model is presented. The Gibbs sampling method via rejection sampling was used in conjunction with the computer program OpenBUGS. Examples that compared the estimation method with another Gibbs sampling method via data augmentation as well as conditional, marginal, and joint maximum likelihood estimation methods are presented using empirical data sets. The effects of prior specifications on the difficulty and ability estimates are illustrated with the empirical data sets. A discussion is presented for related issues of Bayesian estimation in item response theory.

Key Words: Bayesian estimation, data augmentation, Gibbs sampling, rejection sampling, Rasch model.

INTRODUCTION

For the one-parameter logistic Rasch model (Rasch, 1980) many estimation methods can be used to obtain item difficulty and person's ability parameter estimates (Fischer & Molenaar, 1995; Hoijtnik & Boomsma, 1995; Molenaar, 1995). Difficulty and ability parameters can be estimated jointly by maximizing the joint likelihood function (i.e., JML; Wright & Stone, 1979). Conditional maximum likelihood (CML; Andersen, 1980) seems to be the standard estimation method under the one-parameter logistic model for estimation of difficulty parameters (e.g., Molenaar, 1995). Also, marginal maximum likelihood (MML) estimation using the expectation and maximization algorithm can be used to obtain difficulty parameter estimates (du Toit, 2003; Thissen, 1982). In addition, joint Bayesian estimation and marginal Bayesian estimation can be employed to obtain parameter estimates under the one-parameter logistic model (e.g., Birnbaum, 1969; Mislevy, 1986; Swaminathan & Gifford, 1982; see also Tsutakawa, & Lin, 1986).

Point estimates of the Rasch model difficulty and ability parameters are obtained in these earlier maximum likelihood estimation and Bayesian estimation methods by maximizing some forms of the likelihood function or of the posterior distribution. Instead of obtaining point estimates, procedures to approximate the posterior distribution under the Bayesian framework have been proposed relatively recently. One such method, Gibbs sampling approaches the estimation of item and ability parameters using the joint posterior distribution rather than the marginal distribution (e.g., Albert, 1992; Johnson & Albert, 1999; Kim, 2001; Patz & Junker, 1999). It can be noted that there are several different versions and implementations of Gibbs sampling that can be used to estimate item and ability parameters. Even so, all Bayesian estimation methods should yield comparable item and ability

* Asst. Prof. Dr., Harran University, Faculty of Education, Educational Sciences, Şanlıurfa-Turkey, e-mail: sedatsen@harran.edu.tr, ORCID ID: orcid.org/0000-0001-6962-4960

** Asst. Prof. Dr., Kilis 7 Aralık University, Faculty of Education, Educational Sciences, Kilis-Turkey, e-mail: tugba-mat@hotmail.com, ORCID ID: orcid.org/0000-0002-8738-7177

*** Research Prof. Dr. Korea University, Office of Research Management, Seoul, South Korea, e-mail: heom@korea.ac.kr, ORCID ID: orcid.org/0000-0003-0611-1994

**** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: acohen@uga.edu, ORCID ID: orcid.org/0000-0002-8776-9378

***** Prof. Dr., University of Georgia, College of Education, Educational Psychology Department, Athens-Georgia, USA, e-mail: shkim@uga.edu, ORCID ID: orcid.org/0000-0002-2353-7826

DOI: 10.21031/epod.408451
Received: 21.03.2018
Accepted: 16.08.2018

parameter estimates, especially when comparable priors are used or when ignorance or locally-uniform priors are used. This paper was designed to investigate this issue using the one-parameter logistic Rasch model. Specifically, difficulty and ability parameter estimates from a Gibbs sampling method that used the rejection sampling (GS1) is examined and compared with another Gibbs sampling method that used data augmentation (GS2) as well as CML, MML, and JML. Because there exists Swaminathan and Gifford's (1982) seminal paper for Bayesian estimation under the Rasch model, GS1 is explained below with their framework instead of employing new notations. The main issue that differentiates GS1 in the current paper and the implementation used in Swaminathan and Gifford (1982) lies in the notion of the posterior maximization and approximation.

It should be noted that in item response theory Gibbs sampling and the more general Markov chain Monte Carlo methods are originally proposed to estimate parameters in rather complicated item response models for that the usual estimation methods may not be readily available. Although Gibbs sampling and the Markov chain Monte Carlo methods have been successfully applied to the modeling of complex response data in some studies (e.g., Bolt, Cohen, & Wollack, 2001, 2002; Cohen & Bolt, 2005; Karabatsos & Batchelder, 2003; Sen, Cohen, & Kim, 2018) and some specialized computer programs (e.g., Baker, 1998; Johnson & Albert, 1999; Wang, Bradlow, & Wainer, 2005) as well as a general computer program (Spiegelhalter, Thomas, Best, & Gilks, 1997a) have been available, only limited studies are available that investigated the characteristics of parameter estimates from Gibbs sampling or the Markov chain Monte Carlo methods for the traditional item response theory models including the Rasch model. Wollack, Bolt, Cohen, and Lee (2002), for example, investigated the recovery characteristics of Gibbs sampling for the nominal response model, and Baker (1998) investigated the recovery characteristics for the two-parameter logistic model. Kim (2001) reported results from a comparison study for the one-parameter logistic model in which a Gibbs sampling method was contrasted with other maximum likelihood estimation methods. Öztürk and Karabatsos (2017) discussed Gibbs sampling methods for estimating difficulty and ability parameters along with item response outlier detection parameters under the Rasch model. Levy (2009) presented an excellent review of the Markov chain Monte Carlo methods and Gibbs sampling for estimating item response theory models and the discussion of prior specifications for the Bayesian estimation. Interested readers should consult with Levy (2009) and references therein for the various computational methods under the Bayesian framework. Recently, Sheng (2010, 2017) investigated the use or specification of priors on the Markov chain Monte Carlo estimates under the three-parameter normal ogive model. Natesan, Nandakumar, Minka, and Rubright (2016) investigated the effects of priors on the Markov chain Monte Carlo and variational Bayes estimates for the one-, two-, and three-parameter logistic models.

Note that, despite the importance of the specification of priors in Bayesian estimation and the Gibbs sampling method, there is not much transparency regarding the selection and use of priors in the literature. This paper also illustrates the role of priors in the context of hierarchical Bayesian framework of Swaminathan and Gifford (1982) under the Rasch model.

In the subsequent sections, various implementations of the estimation methods for the Rasch model are briefly presented for the maximum likelihood methods and the Bayesian methods with a detailed explication of prior specifications. Results from a comparison study for the various estimation methods for the Rasch model are reported using empirical data from a published article. In order to assess the effects of prior specifications on the parameter estimates in GS1, results from a comparison study for employing various prior specifications are reported. Discussion for the general issues related Bayesian estimation in item response theory is followed.

Implementations of Estimation Methods

Methods of Maximum Likelihood

This paper employed proprietary computer programs for the maximum likelihood estimation of the difficulty and ability parameters. Specifically, WINMIRA (van Davier, 2001) was used for CML, IRTPRO (Cai, Thissen, & du Toit, 2010) was used for MML, and Winsteps (Linacre, 2003) was used

for JML. Technical treatments of these estimation methods can be found in several original articles contained as references in the computer program manuals. Baker and Kim (2004) also contains some accounts of the implementations of the respective methods.

A main reference for CML is Andersen (1980) (see also Andersen, 1970, 1972; Baker & Harwell, 1994). Earlier FORTRAN code of CML can be found in Fischer (1968) and Fischer and Allerup (1968). Thissen (1982) presented detailed accounts for theoretical background and the implementation of MML of difficulty parameters under the Rasch model. The explication of the two versions of Thissen's (1982) MML can be found in Baker and Kim (2004, pp. 397–411) with BASIC and Java code. Wright and his colleagues published many papers that presented implementations of JML (e.g., Wright & Panchapakesan, 1969). FORTRAN code for the earlier predecessors of Winsteps can be found in Wright and Mead (1978) and Wright, Mead, and Bell (1980) (cf. Wright, Linacre, & Schultz, 1989). Although not treated in this manuscript, it should be noted that there are other recent implementations of these earlier methods in R (Venables, Smith, & The R Development Core Team, 2009). Examples of R packages for item response theory modeling include ltm (Rizopoulos, 2006), eRm (Mair & Hatzinger, 2007), and mirt (Chalmers, 2012).

Bayesian Methods

Swaminathan and Gifford (1982) presented Bayesian[1] estimation for the Rasch model. There are other papers that presented Bayesian estimation methods for more general item response theory models (e.g., Leonard & Novick, 1985; Mislevy, 1986; Swaminathan & Gifford, 1985, 1986; Swaminathan, Hambleton, Sireci, Xing, & Rizavi, 2003; Tsutakawa & Lin, 1986). As indicated earlier, nearly all Bayesian methods in item response theory that were implemented on the computer programs were used to obtain parameter estimates by maximizing some form of the posterior distribution.

Only recently, for example, Fox (2010), Stone and Zhu (2015), Levy and Mislevy (2016), and Luo and Jiao (2017) presented Bayesian estimation of item and ability parameters based on the techniques for the approximation of the posterior distribution, although Albert (1992) presented such a method some time ago. Kim and Bolt (2007) presented excellent instructional material for the Markov chain Monte Carlo methods to estimate parameters in item response theory models.

This paper is based on Swaminathan and Gifford's framework and presents its implementation on OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014). It deals with two different Bayesian estimation cases; (1) ability parameter estimation with known difficulty parameters and (2) difficulty and ability parameter estimation. The first case may provide a good foundational information for the second case. These two cases are presented below without employing detailed equations because nearly all of them can be found in Swaminathan and Gifford (1982).

Ability Estimation with Known Difficulty Parameters

In Bayesian ability estimation with known difficulty parameters, the posterior distribution can be defined as

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{p(x)}, \quad (1)$$

where $p(x|\theta)\equiv l(\theta)$ is the likelihood function of the ability parameter θ with item response data x, $p(\theta)$ is the prior distribution, and $p(x) = \int p(x|\theta)p(\theta)d\theta$. Following Lindley and Smith (1972) and

[1]It is not known to us that what will be the Reverend Thomas Bayes's (1701–1761) answer to the question of "Are you a Bayesian? " He was the first by the eponymy to solve the inverse problem of passage from the sample to population using ideas that are very popular today (Dodge, 2003, p. 29; Trader, 1997; cf. Stigler, 1980). Bayes's (1763) original paper was reprinted (see Bayes, 1958) with a biographical note by Barnard (1958). It should be noted that there is a list of eight errata for the original paper (Bayes, 1763) on the supposedly page 543 of the *Philosophical Transactions, Vol. 53.* Barnard's (1958) note didn't indicate that there is the errata page, and the reprint on *Biometrika, Vol. 45* with modern notation did not include two of the errata.

Novick, Lewis, and Jackson (1973), Swaminathan and Gifford (1982) used a hierarchical prior, $p(\theta) = \Pi_i p(\theta_i|\mu,\phi)p(\mu,\phi)$, where i designates each person, $p(\mu,\phi)=p(\phi)$ for which $p(\mu)$ has an improper uniform distribution and $p(\phi)$ has the inverse chi-square distribution with parameters ν and λ (i.e., $\phi \sim \chi^{-2}(\nu,\lambda)$; Novick & Jackson, 1974, pp. 190–194). The nuisance parameters μ and ϕ are integrated out of the posterior distribution and then the resulting proportional posterior distribution is maximized with the Newton-Raphson scheme to obtain point estimates of the ability parameters. With a fixed μ value, the kernel of the resulting ability distribution is that of the multivariate t distribution (Anderson, 1984, pp. 272–273), and all ability parameters are estimated simultaneously in the Newton-Raphson scheme. The specification of the hyperparameters ν and λ is a key issue in such hierarchical Bayesian estimation.

In conjunction with the Markov chain Monte Carlo method for approximating the entire posterior distribution and in the context of the computer program OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014) used in this study, it is better to use a proper yet noninformative uniform or normal hyperprior distribution for μ in addition to employing an independent hyperprior distribution for ϕ. The specification of the hyperparameters for the hyperprior distributions seems to be a very important issue. A noninformative, diffuse hyperprior distribution can be used for μ by specifying appropriate hyperparameters, and an informative hyperprior distribution can be used for ϕ by specifying appropriate hyperparameters.

One problem frequently encountered when specifying the distributional characteristics is that there are too many different definitions of the specific distributions in Bayesian literature (cf. Segal's law; Block, 1977, p. 79). Because this paper is based on Swaminathan and Gifford's notation but uses OpenBUGS to obtain posterior distributional statistics in GS1, it is imperative to connect seemingly the same yet different notations from different sources. An illustration below is for the inverse chi-square distribution and the gamma distribution in essence.

Swaminathan and Gifford (1982, p. 178) used the scaled inverse chi-square distribution for ϕ:

$$p(\phi|\nu,\lambda) \propto \frac{1}{\phi^{\frac{1}{2}\nu+1}} \exp\left[\frac{\lambda}{-2\phi}\right], \quad 0<\phi<\infty, \quad \lambda>0, \quad \nu>0 \tag{2}$$

(see Novick & Jackson, 1974, pp. 190–194; Isaacs, Christ, Novick, & Jackson, 1974, 175–196). Hence $\phi \sim \chi^{-2}(\nu,\lambda)$ and $\phi^{-1} \sim \chi^2(\nu,\lambda^{-1}) = \chi^2(\nu,\omega)$, where $W=\phi^{-1}$ variable has a scaled chi-square density,

$$p(W|\nu,\omega) \propto \frac{W^{(\nu/2)-1}}{\omega^{\nu/2}} \exp\left[\frac{-W}{2\omega}\right], \quad W>0, \quad \nu>0, \quad \omega>0 \tag{3}$$

(see Novick & Jackson, 1974, pp. 186–190). It is not good that functions are shown with proportionality because the exact density of the distribution is not explicit.

In terms of the exact density of the scaled inverse chi-square without employing proportionality (see e.g., Gelman, Carlin, Stern, & Rubin, 1995, pp. 474–475 with their $\theta \equiv \phi$ and $\nu s^2 \equiv \lambda$ of Novick & Jackson, 1974, p. 191),

$$p(\phi|\nu,\lambda) = \frac{(\lambda/2)^{\nu/2}}{\Gamma(\nu/2)} \frac{1}{\phi^{\frac{1}{2}\nu+1}} \exp\left[\frac{\lambda}{-2\phi}\right], \tag{4}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ is a gamma function (Davis, 1964, p. 255). Note that this distribution is Berger's (1985, p. 561) inverse gamma density, $IG(\alpha,\beta)$, where $\alpha=\nu/2$ and $\beta=2/\lambda$ (n.b., this β is not the difficulty parameter).

In prior specification, a different but better form of the distribution can be used. If $\nu\lambda_l/\phi \sim \chi^2(\nu)$ (Lindley, 1965, p. 26; Leonard & Hsu, 1999, p. 214; subscript *l* designates λ from Lindley and Leonard & Hsu), then

$$p(\phi|\nu,\lambda_l)=\frac{(\nu\lambda_l/2)^{\nu/2}}{\Gamma(\nu/2)}\frac{1}{\phi^{\frac{1}{2}\nu+1}}\exp\left[-\frac{\nu\lambda_l}{2\phi}\right], \quad (5)$$

where ν is the prior sample size and λ_l^{-1} is the prior mean of ϕ^{-1} with the prior mean of φ to be $\nu\lambda_l/(\nu-2)$ for ν>2. In terms of Berger's *IG*(α,β), the corresponding parameters should be α=ν/2 and β=2/(νλ$_l$). In terms of Swaminathan and Gifford's (1982, p. 178) $\chi^{-2}(\nu,\lambda)$, ν=ν and λ=νλ$_l$ of Lindley (1965, p. 26), yielding the prior sample size is ν, the prior mean of ϕ^{-1} is ν/λ, and the prior mean of φ is λ/(ν−2) for ν>2.

These distributions may not be directly used in available computer software. In OpenBUGS, WinBUGS, as well as BUGS (e.g., Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013, pp. 345–346), φ~dgamma(*a*,*b*) denotes the density is

$$p(\phi|a,b)=b^a\phi^{a-1}e^{-b\phi}/\Gamma(a) \quad \text{for} \quad \phi>0,\ a,b>0 \quad (6)$$

with mean *a*/*b* and variance *a*/*b*^{2}. In Berger's (1985, p. 560) gamma density, *G*(α,β), the parameters are α=*a* and β=1/*b* with mean αβ and variance αβ^2. Note that φ~*IG*(ν/2,2/λ) means

$\phi^{-1} \sim G(\nu/2,2/\lambda) = \text{dgamma}(\nu/2,2/\lambda)$ in OpenBUGS with ν=2*a* to be the prior sample size, ν/λ=*a*/*b* to be the prior mean of ϕ^{-1}, and λ/(ν−2)=*b*/(*a*−1) to be the prior mean of φ for ν=2*a*>2.

Estimation of Both Difficulty and Ability Parameters

The posterior distribution in this case can be defined as

$$p(\theta,\beta|x)=\frac{p(x|\theta,\beta)p(\theta,\beta)}{p(x)}, \quad (7)$$

where $p(x|\theta,\beta)\equiv l(\theta,\beta)$ is the likelihood function of the ability parameter θ and the difficulty parameter β with item response data *x*, $p(\theta,\beta)$ is the prior distribution, and $p(x)=\int p(x|\theta,\beta)p(\theta,\beta)d(\theta,\beta)$. Again, following Lindley and Smith (1972) and Novick, Lewis, and Jackson (1973), Swaminathan and Gifford (1982) used independent hierarchical priors, $p(\theta,\beta)=p(\theta)p(\beta)=\Pi_i p(\theta_i|\mu_\theta,\phi_\theta)p(\mu_\theta,\phi_\theta)\times\Pi_j p(\beta_j|\mu_\beta,\phi_\beta)p(\mu_\beta,\phi_\beta)$, where *i* designates each person and *j* designates each item, $p(\mu_\theta,\phi_\theta)=p(\phi_\theta)$ and $p(\mu_\beta,\phi_\beta)=p(\phi_\beta)$ for which $p(\mu_\theta)$ and $p(\mu_\beta)$ have improper uniform distributions and $p(\phi_\theta)$ and $p(\phi_\beta)$ have the inverse chi-square distributions with parameters ν_θ, λ_θ, ν_β, λ_β, respectively (i.e., $\phi_\theta \sim \chi^{-2}(\nu_\theta,\lambda_\theta)$ and $\phi_\beta \sim \chi^{-2}(\nu_\beta,\lambda_\beta)$). Again, the nuisance parameters μ_θ, ϕ_θ, μ_β, ϕ_β are integrated out of the posterior distribution and then the resulting proportional posterior distribution is maximized with the Newton-Raphson scheme to obtain point estimates of the ability and item parameters. An iterative Birnbaum paradigm is used to obtain a set of ability estimates and then a set of difficulty parameter estimates until the overall convergence criterion can be met (Swaminathan & Gifford, 1982, p. 184).

The specification of the hyperparameters (i.e., ν_θ, λ_θ, ν_β, λ_β) is a key issue in hierarchical Bayesian estimation. In conjunction with the Markov chain Monte Carlo method for approximating the entire

posterior distribution and in the context of the computer program OpenBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2014) used in this study, it is better to use a proper yet noninformative uniform or normal hyperprior distribution for μ_θ or μ_β in addition to employ an independent hyperprior for ϕ_θ or ϕ_β. A noninformative, diffuse hyperprior distribution can be used for each μ by specifying appropriate hyperparameters, and an informative hyperprior distribution can be used for each ϕ by specifying appropriate hyperparameters.

METHOD

Without loss of generality, we present below a comparison study for estimation of both difficulty and ability parameters under Rasch model. Ability estimation can also be done by modifying the programs in a trivial manner and hence not presented.

To compare GS1, GS2, CML, MML, and JML, illustrations using (1) the Law School Admission Test-Section 6 (LSAT6; Bock & Aitkin, 1981; Bock & Lieberman, 1970) data and (2) the Law School Admission Test-Section 7 (LSAT7) are presented below. It should be noted that the LSAT6 and LSAT7 data have been analyzed in many published articles and books (e.g., Andersen, 1980; McDonald, 1999). Use of these data instead of employing simulation data, hence, may provide a familiar baseline to make comparisons of different estimation methods.

GS1 estimates were obtained using OpenBUGS. GS2 estimates were obtained using MATLAB (The MathWorks, 1996) employing the code from Johnson and Albert (1999). Instead of OpenBUGS, WinBUGS or BUGS (e.g., Spiegelhalter et al., 1997a) can also be used. Difficulty parameter estimates are reported first and ability parameter estimates are subsequently reported for LSAT6 and LSAT7, respectively. It is not necessary to show the listings of the input lines of CML, MML, and JML. Also for GS2, the MATLAB function presented in Johnson and Albert (1999, p. 248) was used without any modification. However, it is necessary to present the input lines for OpenBUGS. The portions of the input lines are contained in Appendix. Note that in Appendix the inverse of the hyperparameter variance was specified with dgamma (a=2.5, b=5) for both ability and difficulty prior distributions. This prior specification is equivalent to Swaminathan and Gifford's (1982) ν=5 and λ=10. Also note that the centered value of the log odds of the classical item facilities denoted as p_j (i.e., values of $\log[(1-p_j)/p_j]$ centered at 0) were used for the initial values for difficulty parameters. Similar initial values were specified for the ability parameters.

Based on the suggestions from Kim and Bolt (2007) and Kim (2001), burn-in was set to 1000 and the next 10,000 iterations were used for GS1 to construct the posterior distributions that showed convergence of the simulated draws (see Gilks, Richardson, & Spiegelhalter, 1996). The convergence of the chains was visually monitored by checking history and autocorrelation plots. It should be noted that there are many different ways to summarize the sampled values in GS1 or GS2. Instead of using the actual posterior credibility interval, the posterior means and the posterior standard deviations are used in this study. The marginal posterior densities of the samples values for respective parameters all followed unimodal and likely normal distributions in GS1. GS2 also yielded similar results for the sampled values.

RESULTS

Comparison of Estimation Methods

LSAT6 Estimation Results

For the LSAT6 data that contained responses of 1000 subjects to five items, all five methods yielded practically the same results for the difficulty estimates. Table 1 presents difficulty parameter estimates based on the usual Rasch model scaling (i.e., the mean of difficulties is zero) that is the default setting

for nearly all Rasch model calibration computer programs. Note that some differences still exist among the difficulty parameter estimates and the accompanied standard errors or posterior standard deviations. Although results from this simple data set may not be sufficient for fully evaluating different estimation methods, these may provide good enough information about the agreement in estimation results.

Table 1. LSAT6 Difficulty Estimates

	GS1		GS2[a]		CML		MML[a]		JML	
Item	b_j	(p.s.d.)	b_j	(p.s.d.)	b_j	(s.e.)	b_j	(s.e.)	b_j	(s.e.)
1	–1.26	(0.11)	–1.38	(0.10)	–1.26	(0.13)	–1.26	(0.13)	–1.24	(0.11)
2	0.48	(0.07)	0.52	(0.07)	0.47	(0.08)	0.48	(0.08)	0.45	(0.07)
3	1.25	(0.07)	1.43	(0.07)	1.24	(0.08)	1.24	(0.07)	1.30	(0.07)
4	0.17	(0.07)	0.16	(0.08)	0.17	(0.09)	0.17	(0.09)	0.13	(0.07)
5	–0.63	(0.09)	–0.72	(0.09)	–0.62	(0.11)	–0.63	(0.11)	–0.64	(0.08)

Note. p.s.d. = posterior standard deviation; s.e. = standard error

[a]Estimates were transformed onto the zero centered logistic metric.

LSAT6 ability estimates and either the accompanied standard errors or the posterior standard deviations are reported in Table 2 for each number-correct raw score from 0 to 5. In GS1 and GS2 there were different posterior means for examinees with the same response pattern or the same raw score. In reporting of the ability estimates, the first examinees who got the respective raw scores were used to obtain the estimates (i.e., examinees 1, 4, 12, 28, 62, and 703). Although the estimates who got the same raw score were trivially different in the consideration of the magnitude of the posterior standard deviation, obtaining such odd results were not seen in other maximum likelihood based estimation procedures.

The most pronounced pattern in Table 2 is that estimates from GS1 and MML/EAP (i.e., expected a posteriori) were very similar. Other estimation methods look somewhat different due to the extremely small test size. Except for the scores 0 and 5, however, ability estimates from CML/ML and JML were very similar. Because in the Rasch model with conditional maximum likelihood estimation the weighted likelihood estimation (WLE; Warm, 1989) is popular, the results for such a case were reported in the CML/WLE column.

Table 2. LSAT6 Ability Estimates

	GS1		GS2[a]		CML/ML		CML/WLE		MML/EAP[a]		JML[b]	
Score	θ_i	(p.s.d.)	θ_i	(p.s.d.)	θ_i	(s.e.)	θ_i	(s.e.)	θ_i	(p.s.d.)	θ_i	(s.e.)
0	–0.09	(0.64)	–1.61	(0.98)			–2.79	(1.72)	0.03	(1.05)	–3.22	(1.93)
1	0.31	(0.64)	–0.74	(0.91)	–1.60	(1.18)	–1.34	(1.11)	0.40	(1.05)	–1.72	(1.21)
2	0.71	(0.64)	0.02	(0.87)	–0.47	(0.99)	–0.41	(0.99)	0.76	(1.07)	–0.52	(1.03)
3	1.12	(0.66)	0.79	(0.85)	0.48	(0.99)	0.42	(0.98)	1.14	(1.11)	0.51	(1.21)
4	1.56	(0.67)	1.48	(0.91)	1.60	(1.18)	1.34	(1.11)	1.54	(1.11)	1.72	(1.21)
5	2.02	(0.70)	3.32	(1.24)			2.78	(1.71)	1.95	(1.13)	3.28	(1.93)

Note. p.s.d. = posterior standard deviation; s.e. = standard error. GS1 and GS2 estimates were from examinees 1, 4, 12, 28, 62, and 703.

[a]Estimates were transformed onto the zero centered logistic metric of item difficulty.

[b]Ad hoc estimates were inserted to scores 0 and 5, respectively.

LSAT7 Estimation Results

For the LSAT7 data, all five methods yielded practically the same results for the difficulty estimates as did for the LSAT6 data. Table 3 presents difficulty parameter estimates based on the usual Rasch model scaling. Note that some differences still exist among the difficulty parameter estimates and the accompanied standard errors or posterior standard deviations.

Table 3. LSAT7 Difficulty Estimates

	GS1	GS2[a]	CML	MML[a]	JML
Item	b_j (p.s.d.)	b_j (p.s.d.)	b_j (s.e.)	b_j (s.e.)	b_j (s.e.)
1	–0.54 (0.08)	–0.59 (0.14)	–0.54 (0.10)	–0.54 (0.13)	–0.55 (0.08)
2	0.54 (0.07)	0.59 (0.12)	0.54 (0.08)	0.54 (0.09)	0.53 (0.07)
3	–0.13 (0.07)	–0.17 (0.14)	–0.13 (0.09)	–0.13 (0.11)	–0.15 (0.07)
4	0.81 (0.07)	0.90 (0.11)	0.81 (0.08)	0.80 (0.09)	0.83 (0.07)
5	–0.67 (0.08)	–0.73 (0.15)	–0.67 (0.10)	–0.66 (0.14)	–0.67 (0.08)

Note. p.s.d. = posterior standard deviation; s.e. = standard error

[a]Estimates were transformed onto the zero centered logistic metric.

Table 4 shows the ability estimates and either the accompanied standard errors or the posterior standard deviations for each number-correct raw score from 0 to 5 for LSAT7. As was the case for LSAT6, in GS1 and GS2 there were different posterior means for examinees with the same response pattern or the same raw score. In reporting of the ability estimates, the first examinees who got the respective raw scores were used to obtain the estimates (i.e., examinees 1, 13, 33, 65, 145, and 693).

Note that ability estimates from GS1 and MML/EAP were very similar in Table 4. Other estimation methods yielded somewhat different ability estimates partly due to the extremely small test size. Except for the scores 0 and 5, however, ability estimates from CML/ML and JML were very similar.

Table 4. LSAT7 Ability Estimates

	GS1	GS2[a]	CML/ML	CML/WLE	MML/EAP[a]	JML[b]
Score	θ_i (p.s.d.)	θ_i (p.s.d.)	θ_i (s.e.)	θ_i (s.e.)	θ_i (p.s.d.)	θ_i (s.e.)
0	–0.63 (0.73)	–1.72 (1.00)		–2.57 (1.66)	–0.59 (0.70)	–2.96 (1.90)
1	–0.12 (0.71)	–0.81 (0.91)	–1.49 (1.14)	–1.21 (1.07)	–0.10 (0.69)	–1.54 (1.16)
2	0.38 (0.72)	0.11 (0.90)	–0.44 (0.95)	–0.38 (0.94)	0.39 (0.70)	–0.47 (0.97)
3	0.91 (0.73)	0.78 (0.91)	0.44 (0.95)	0.37 (0.95)	0.89 (0.72)	0.45 (0.97)
4	1.47 (0.77)	1.54 (0.94)	1.49 (1.15)	1.21 (1.07)	1.44 (0.75)	1.54 (1.16)
5	2.11 (0.83)	2.86 (1.16)		2.59 (1.67)	2.05 (0.80)	2.98 (1.91)

Note. p.s.d. = posterior standard deviation; s.e. = standard error. GS1 and GS2 estimates were from examinees 1, 13, 33, 65, 145, and 693.

[a]Estimates were transformed onto the zero centered logistic metric of item difficulty.

[b]Ad hoc estimates were inserted to scores 0 and 5, respectively.

Comparison of Prior Specifications

To assess the effects of prior specifications on the difficulty and ability parameter estimates, the same LSAT6 and LSAT7 data were analyzed with OpenBUGS. Four prior specifications with four different sets of hyperparameters were used for both ability and difficulty prior distributions; (1) dgamma(a=2.5, b=5), (2) dgamma(a=4, b=5), (3) dgamma(a=7.5, b=5), and (4) dgamma(a=12.5, b=5). Because the first specification was the same as in the earlier calibration condition, only three additional OpenBUGS runs were performed for LSAT6 and LSAT7, respectively. Except for the prior specification, all other settings to obtain the estimates remained the same for the OpenBUGS runs.

Note that these prior specifications of a=2.5, 4, 7.5, 12.5 with b=5 are fully equivalent to Swaminathan and Gifford's (1982) ν=5, 8, 15, 25 with λ=10 used in their study.

LSAT6 Prior Specification Results

For the LSAT6 data, all four prior specifications yielded practically the same results for the difficulty estimates, but a bit different results for the ability estimates. Table 5 presents difficulty parameter estimates based on the usual Rasch model scaling. Note that only trivial differences exist among the difficulty parameter estimates and the posterior standard deviations, that occur in the second decimal places. Because each difficulty parameter was estimated with the sample size of 1000, shrinkage toward the mean of the difficulty estimates might exist with the increasing hyperparameter a values but barely noticeable. In Figure 1(a) LSAT6 difficulty estimates are plotted with the four different values of the hyperparameter a=2.5, 4, 7.5, 12.5 (because the hyperparameter b=5 for all cases only the four hyperparameters of a were used). The numbers in the plot designate the item numbers.

Table 5. LSAT6 Difficulty Estimates from Prior Specifications

	GS1 Hyperparameters							
	a=2.5, b=5		a=4, b=5		a=7.5, b=5		a=12.5, b=5	
Item	b_j	(p.s.d.)	b_j	(p.s.d.)	b_j	(p.s.d.)	b_j	(p.s.d.)
1	–1.26	(0.11)	–1.25	(0.10)	–1.24	(0.10)	–1.22	(0.10)
2	0.48	(0.07)	0.48	(0.07)	0.47	(0.07)	0.46	(0.07)
3	1.25	(0.07)	1.24	(0.07)	1.23	(0.07)	1.21	(0.07)
4	0.17	(0.07)	0.17	(0.07)	0.16	(0.07)	0.16	(0.07)
5	–0.63	(0.09)	–0.63	(0.08)	–0.62	(0.08)	–0.61	(0.08)

Note. p.s.d. = posterior standard deviation

LSAT6 ability estimates from the four prior specifications and the posterior standard deviations are reported in Table 6 for each number-correct raw score from 0 to 5. In GS1 there were different posterior means for examinees with the same response pattern or the same raw score. In reporting of the ability estimates, the first examinees who got the respective raw scores were used to obtain the estimates (i.e., examinees 1, 4, 12, 28, 62, and 703).

Considering the magnitude of the posterior standard deviations, it can be noted in Table 6 that practically trivial differences exist among the ability estimates and the posterior standard deviations. Nevertheless, because each ability parameter was estimated with the truly small number of items, shrinkage toward the mean of ability estimates with the increasing hyperparameter a values was quite noticeable. In Figure 1(b) LSAT6 ability estimates are plotted with the four different values of the hyperparameter a=2.5, 4, 7.5, 12.5 (because the hyperparameter b=5 for all cases only the four hyperparameters of a were used). The numbers in the plot designate the raw scores from 0 to 5.

Table 6. LSAT6 Ability Estimates from Four Prior Specifications

	GS1 Hyperparameters							
	a=2.5, b=5		a=4, b=5		a=7.5, b=5		a=12.5, b=5	
Score	θ_i	(p.s.d.)	θ_i	(p.s.d.)	θ_i	(p.s.d.)	θ_i	(p.s.d.)
0	–0.09	(0.64)	–0.07	(0.64)	0.01	(0.62)	0.12	(0.60)
1	0.31	(0.64)	0.33	(0.63)	0.39	(0.62)	0.45	(0.59)
2	0.71	(0.64)	0.72	(0.64)	0.77	(0.62)	0.79	(0.59)
3	1.12	(0.66)	1.13	(0.64)	1.13	(0.62)	1.15	(0.60)
4	1.56	(0.67)	1.56	(0.65)	1.54	(0.64)	1.50	(0.61)
5	2.02	(0.70)	2.02	(0.68)	1.97	(0.66)	1.89	(0.63)

Note. p.s.d. = posterior standard deviation

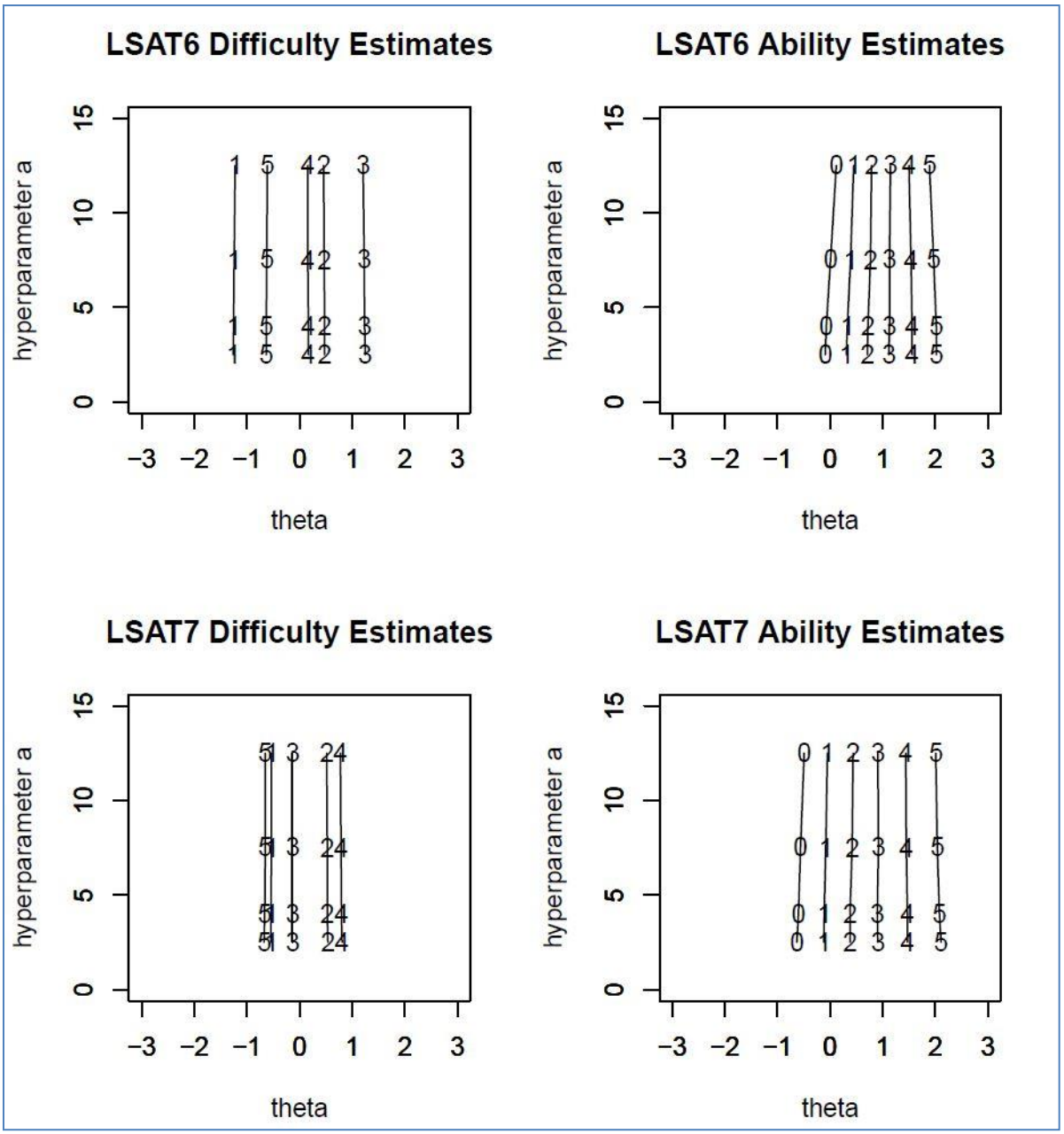


Figure 1. Plots of (a) LSAT6 difficulty estimates, (b) LSAT6 ability estimates, (c) LSAT7 difficulty estimates, and (d) LSAT7 ability estimates for the hyperparameter values of a=2.5, 4, 7.5, 12.5 with b=5.

LSAT7 Prior Specification Results

For the LSAT7 data, all four prior specifications yielded practically the same results for the difficulty estimates, but a bit different results for the ability estimates. Table 7 presents difficulty parameter estimates based on the usual Rasch model scaling. Note that only trivial differences exist among the difficulty parameter estimates and the posterior standard deviations, that occur in the second decimal places. Because each difficulty parameter was estimated with the sample size of 1000, shrinkage toward the mean of difficulty estimates might exist but not really noticeable. In Figure 1(c) LSAT7

difficulty estimates are plotted with the four different values of the hyperparameter *a*=2.5, 4, 7.5, 12.5. The numbers in the plot designate the item numbers.

Table 7. LSAT7 Item Difficulty Estimates from Four Prior Specifications

	GS1 Hyperparameters							
	a=2.5, *b*=5		*a*=4, *b*=5		*a*=7.5, *b*=5		*a*=12.5, *b*=5	
Item	b_j	(p.s.d.)	b_j	(p.s.d.)	b_j	(p.s.d.)	b_j	(p.s.d.)
1	–0.54	(0.08)	–0.54	(0.08)	–0.53	(0.08)	–0.53	(0.08)
2	0.54	(0.07)	0.53	(0.07)	0.53	(0.07)	0.52	(0.07)
3	–0.13	(0.07)	–0.13	(0.07)	–0.13	(0.07)	–0.13	(0.07)
4	0.81	(0.07)	0.80	(0.07)	0.79	(0.07)	0.78	(0.07)
5	–0.67	(0.08)	–0.66	(0.08)	–0.65	(0.08)	–0.65	(0.08)

Note. p.s.d. = posterior standard deviation

LSAT7 ability estimates from the four prior specifications and the posterior standard deviations are reported in Table 8 for each number-correct raw score from 0 to 5. In GS1 there were different posterior means for examinees with the same response pattern or the same raw score. In reporting of the ability estimates, the first examinees who got the respective raw scores were used to obtain the estimates (i.e., examinees 1, 13, 33, 65, 145, and 693).

It can be noted that practically trivial differences exist among the ability estimates and the posterior standard deviations, considering the magnitude of the posterior standard deviations. Nevertheless, each ability parameter was estimated with the truly small number of items, shrinkage toward the mean of ability estimates with the increasing hyperparameter *a* values was quite noticeable. In Figure 1(d) LSAT7 ability estimates are plotted with the four different values of the hyperparameter *a*=2.5, 4, 7.5, 12.5. The numbers in the plot designate the raw scores from 0 to 5.

Table 8. LSAT7 Ability Estimates from Four Prior Specifications

	GS1 Hyperparameters							
	a=2.5, *b*=5		*a*=4, *b*=5		*a*=7.5, *b*=5		*a*=12.5, *b*=5	
Score	θ_i	(p.s.d.)	θ_i	(p.s.d.)	θ_i	(p.s.d.)	θ_i	(p.s.d.)
0	–0.63	(0.73)	–0.60	(0.73)	–0.56	(0.71)	–0.49	(0.69)
1	–0.12	(0.71)	–0.11	(0.72)	–0.08	(0.69)	–0.05	(0.69)
2	0.38	(0.72)	0.38	(0.72)	0.42	(0.69)	0.44	(0.70)
3	0.91	(0.73)	0.90	(0.73)	0.92	(0.72)	0.91	(0.71)
4	1.47	(0.77)	1.47	(0.77)	1.45	(0.75)	1.44	(0.73)
5	2.11	(0.83)	2.08	(0.83)	2.04	(0.80)	2.01	(0.78)

Note. p.s.d. = posterior standard deviation

DISCUSSION and CONCLUSION

The main difference between the two Gibbs sampling methods, GS1 and GS2, lies in both the specifications of prior distributions and the underlying sampling procedures. The prior distributions used in GS1 had the hierarchical form following Swaminathan and Gifford (1982). For example, the hyperparameter mean of the normal prior distribution for ability had a noninformative uniform distribution and the inverse of the hyperparameter variance of the normal prior had a gamma distribution. In GS1 with gamma(a=2.5, b=5) the prior sample size of the gamma distribution was specified as 2(2.5)=5 and the prior expected value was 2.5/5=0.5 (i.e., the expected value of the hyperparameter variance to be 5/1.5=3.33). Note that this prior specification is equivalent to Swaminathan and Gifford's (1982) ν=5 and λ=10, one of the prior specifications in their paper. They

used three other prior specifications that were converted to the equivalent specifications in the second study. The use of gamma(2.5, 5) seems reasonable among the choices. Swaminathan and Gifford (1982) concluded similarly. Note that there are also other ways of specifying priors for the Rasch model (see Kim, 2001; Levy & Mislevy, 2016; Spiegelhalter et al., 1997b; Stone & Zhu, 2015) instead of using priors in the hierarchical form. In Johnson and Albert's (1999) item_r1 function for GS2 the hyperparamaters of the theta prior was set to have a standard normal distribution while prior standard deviation of the item difficulty parameters was set to unity. See Johnson and Albert (1999, pp. 202–204) for the detailed Gibbs sampling for GS2. Hence GS1 and GS2 differ not only the mathematical forms of the model but also the priors employed.

Because the full conditional distributions for the Rasch model are log-concave (Ghosh, Ghosh, Chen, & Agresti, 1999), the sampling in GS1 used the derivative-free adaptive rejection sampling algorithm (Gilks, 1996; Gilks & Wild, 1992). Due to the use of hierarchical prior distributions, more general sampling procedures can be employed for various parameters in GS1 (see Lunn et al., 2013, pp. 68–70) that include slice sampling (Neal, 2003) and Metropolis-within-Gibbs (Metropolis et al., 1953; Hasting, 1970). In GS2, direct Gibbs sampling method was used with data augmentation because the actual item response theory model was that of the normal ogive instead of the logistic ogive (Albert, 1992; Baker, 1998). The resulting parameter estimates in GS2 were initially expressed on the normal ogive metric but placed onto the logistic metric.

When difficulty and ability are estimated together in GS1 or GS2, the ability estimate for specific case is not unique. The same response pattern may yield different ability estimates and that is not acceptable in practice. In addition, because of employing the exchangeability concept, all ability estimates are estimated simultaneously and there exists some dependency in the resulting estimates. Although estimates are not independent in general, it seems troublesome that estimating ability even with known item parameters may yield different estimates for a specific response pattern. Hence, Gibbs sampling methods or some other estimation methods based on Markov chain Monte Carlo may not be seen as viable methods for the usual item and ability parameter estimation for the usual item response theory models for dichotomous items that include the Rasch model.

In this study, the Rasch model was employed without addressing the problem of model selection, choice of link function, or model fit. Kim and Bolt (2007) contains an excellent introductory review of these issues. Interested readers should refer to Kim and Bolt (2007) and other general references including Lunn et al. (2013).

Note that although Gibbs sampling methods and some computer programs which implemented such procedures have been available sometime, the accuracy of the methods has not been thoroughly studied. Obviously these techniques have been applied to some complicated modeling situations where the traditional maximum likelihood based methods are too difficult to implement, and hence have not been thoroughly tested and compared. Because maximum likelihood based methods have not been implemented at all in such applications, still we need to investigate the relevant estimation procedures. In addition, because there are many different ways of implementing Gibbs sampling methods in item response theory and many different prior distributions can be employed with many different specifications in Bayesian estimation, the illustrative implementation of the Gibbs sampling method and comparing results with other existing Bayesian and likelihood based methods should provide measurement specialists and test developers as well as the users of the computer programs with guidelines for using the Gibbs sampling method under the Rasch item response theory model.

In this study, explications of nearly all estimation methods for the Rasch model were presented together with the two methods based on Gibbs sampling. The specification of priors for ability and difficulty parameters in Bayesian estimation and the Gibbs sampling method was fully explained with detailed mathematical statistical formulas, basically following the framework of Swaminathan and Gifford (1982). Illustrations about the effects of prior specifications on the estimates were presented with empirical data. It should be noted that additional, full scale simulation studies as well as more cumulative experience with regard to prior specifications for Bayesian estimation are definitely needed.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society, Series B, 32,* 283–301.

Andersen, E. B. (1972). The numerical solution to a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 34,* 42–54.

Andersen, E. B. (1980). *Discrete statistical models with social science applications.* Amsterdam,: North-Holland.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling approach. *Applied Psychological Measurement, 22,* 153–169.

Baker, F. B., & Harwell, M. R. (1994). *Estimation of item parameters in the Rasch model via conditional maximum likelihood: A didactic.* Unpublished manuscript, Department of Educational Psychology, University of Wisconsin, Madison, WI.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Barnard, G. A. (1958). Thomas Bayes - a biographical note. *Biometrika, 45,* 293–295.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (of the Royal Society of London), 53,* 370-418; Errara, c. 543.

Bayes, T. (1958). An essay towards solving a problem in the doctrine of changes. *Biometrika, 45,* 296–315.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6,* 258–276.

Block, A. (1977). *Murphy's law and other reasons why things go wrong!* Los Angeles, CA: Price/Stern/Sloan.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics, 26,* 381–409.

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.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Applications of an EM algorithm. *Psychometrika, 46,* 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Cai, L., Thissen, D., & du Toit, S. (2010). *IRTPRO: Item response theory for patient-reported outcomes* [Computer software]. Skokie, IL: Scientific Software International.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29.

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

Davis, P. J. (1964). Gamma function and related functions. In M. Abramowitz & I. A. Stegun (Eds.), *Handbook of mathematical functions with formulas, graphs, and mathematical tables* (pp. 253–293). Washington, DC: National Bureau of Standards.

Dodge, Y. (Ed.). (2003). *The Oxford dictionary of statistical terms.* Oxford, Great Britain: Oxford University Press.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Chicago, IL: Scientific Software International.

Fischer, G. H. (1968). *Einführung in die theorie psychologischer tests: Grunddlagen und anwendungen [Introduction to the theory of psychological tests: Foundations and applications]*. Bern, Switzerland: Huber.

Fischer, G. H., & Allerup, P. (1968). Rechentechnische fragen zu raschs eindimensionalem modell [Computational questions on Rasch's unidimensional model]. In G. H. Fischer (Hrsg. [Ed.]), *Psychologsche Testtheorie [Psychological test theory]* (pp. 269–280). Bern, Switzerland: Huber.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications.* New York, NY: Springer-Verlag.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* New York, NY: Springer.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London, Great Britain: Chapman & Hall.

Ghosh, M., Ghosh, A., Chen, M.-H., & Agresti, A. (1999). *Bayesian estimation for item response model* (Tech. Rep.). Gainsville, FL: University of Florida, Department of Statistics.

Gilks, W. R. (1996). Full conditional distribution. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 75–88). London, England: Chapman and Hall.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* London, England: Chapman and Hall.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics, 41,* 337–348.

Hasting, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57,* 97–109.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 53–68). New York, NY: Springer-Verlag.

Isaacs, G. I., Christ, D. E., Novick, M. R., & Jackson, P. H. (1974). *Tables for Bayesian statisticians.* The Iowa Testing Program, The University of Iowa, Iowa City, IA.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling.* New York, NY: Springer.

Karabatsos, G., & Batchelder, W. H. (2003). Markov chain Monte Carlo estimation for test theory without an answer key. *Psychometrika, 68,* 373–389.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38–51.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163–176.

Leonard, T., & Hsu, J. S. J. (1999). *Bayesian methods: An analysis for statisticians and interdisciplinary researchers.* New York, NY: Cambridge University Press.

Leonard, T., & Novick, M. R. (1985). *Bayesian inference and diagnostics for the three parameter logistic model* (ONR Technical Report No. 85-5). Iowa City, IA: The University of Iowa, Cada Research Group. (ERIC Document Reproduction Service No. ED261068)

Levy, R. (2009). The rise of Markov chain Monte Carlo estimation for psychometric modeling. *Journal of Probability and Statistics, 2009,* ID 537139. doi:10.155/2007/537139

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling.* Boca Raton, FL: CRC Press.

Linacre, J. M. (2003). WINSTEPS Rasch measurement computer program [Computer software]. Chicago, IL: Winsteps.com.

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian viewpoint: Part 2, Inference.* London, Great Britain: Cambridge University Press.

Lindley, D. V., & Smith, A. F. (1972). Bayesian estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1–41.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book: A practical introduction the Bayesian analysis.* Boca Raton, FL: CRC Press.

Luo, Y., & Jiao, H. (2017). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement,* Advance Online Publication. https://doi.org/10.1177/0013164417693666

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21,* 1087–1092.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 39–51). New York, NY: Springer-Verlag.

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in Psychology, 7:1422*. doi:10.3389/fpsyg.2016.01422

Neal, R. M. (2003). Slice sampling. *Annals of Statistics, 31,* 705–741.

Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research.* New York, NY: McGraw-Hill.

Novick, M. R., Lewis, C., & Jackson, P. H. (1973). The estimation of proportions in *n* groups. *Psychometrika, 38,* 19–46.

Öztürk, N., & Karabatsos, G. (2017). A Bayesian robust IRT outlier detecion model. *Applied Psychological Measurement, 41,* 195–208.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Rasch, G. (1980). *Probabilistic model for some intelligence and attainment tests* (With a foreword and afterword by B. D. Wright). Chicago, IL: The University of Chicago Press.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software, 17*(5), 1–25.

Sen, S., Cohen, A. S., & Kim, S. H. (2018). Model selection for multilevel mixture Rasch models. *Applied Psychological Measurement*, 1-18. doi: 10.1177/0146621618779990

Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika, 37*, 87–110.

Sheng, Y. (2017). Investigating a weakly informative prior for item scale hyperparameters in hierarchical 3PNO IRT models. *Frontiers in Psychology, 8:123*. doi:10.3389/fpsyg.2017.00123

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1997a). *BUGS: Bayesian inference using Gibbs sampling* (Version 0.6) [Computer software]. Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1997b). *BUGS 0.5 examples* (Vol. 1, Version i). Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2014). *OpenBUGS user manual.* Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Stigler, S. M. (1980). Stigler's law of eponymy. *Transactions of the New York Academy of Sciences, Series 2, 39*, 147–157.

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Cary, NC: SAS Institute.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175–191.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349–364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589–601.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*, 27–51.

The MathWorks. (1996). MATLAB (Version 5) [Computer program]. Natick, MA: Author.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*, 175–186.

Trader, R. L. (1997). Bayes, Thomas. in N. L. Johnson & S. Kotz (Eds.), *Leading personalities in statistical sciences: From the seventeenth century to the present* (pp. 11–14). New York, NY: John Wiley & Sons.

Tsutakawa, R. K., & Lin, H. Y. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251–267.

Venables, W. N., Smith, D. M., & The R Development Core Team. (2009). *An introduction to R* (2nd ed.). La Vergne, TN: Network Theory.

von Davier, M. (2001). WINMIRA 2001 [Computer program]. St. Paul, MN: Assessment Systems Corporation.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). *User's guide for SCORIGHT (Version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis* (ETS Research Rep. RR-04-49). Princeton, NJ: Educational Testing Service.

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

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339–352.

Wright, B. D., & Linacre, J. M., & Schultz, M. (1989). *A user's guide to BIGSCALE: Rasch-model rating scale analysis computer program.* Chicago, IL: MESA Press.

Wright, B. D., & Mead, R. J. (1978). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23A). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., Mead, R. J., & Bell, S. R. (1980). *BICAL: Calibrating items with the Rasch model* (Research Memorandum No. 23C). Chicago, IL: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23–48.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago, IL: MESA Press.

Rasch Modelinde Gibbs Örnekleme Yönteminin Uygulanması

Giriş

Tek parametreli lojistik Rasch modelinde (Rasch, 1980), madde güçlüğü ve kişi yetenek parametre kestirimlerini elde etmek için birçok kestirim metodu kullanılabilir (Fischer ve Molenaar, 1995; Molenaar, 1995; Hoijtnik ve Boomsma, 1995). Madde güçlük ve kişi yetenek parametreleri, ortak olabilirlik fonksiyonunu maksimize ederek ortak olarak kestirilebilir (yani, JML; Wright ve Stone, 1979). Koşullu maksimum olabilirlik (CML; Andersen, 1980), madde güçlük parametrelerinin tahmini için tek parametreli lojistik modelin altında standart kestirim metodu olarak görünmektedir (ör. Molenaar, 1995). Ayrıca, beklenti ve maksimizasyon algoritmasını kullanarak marjinal maksimum olabilirlik (MML) kestirimi, madde güçlük parametre kestirimlerini elde etmek için kullanılabilir (du Toit, 2003; Thissen, 1982). Ek olarak, tek parametreli lojistik model altında parametre kestirimlerini elde etmek için ortak Bayes kestirimi ve marjinal Bayes kestirimi kullanılabilir (ör. Birnbaum, 1969; Mislevy, 1986; Swaminathan & Gifford, 1982; ayrıca bkz. Tsutakawa, & Lin, 1986).

Rasch modeli madde güçlük ve kişi yetenek parametrelerinin nokta tahminleri, bu olasılık fonksiyonlarını veya sonsal (posterior) dağılımın bazı formlarını maksimize ederek, maksimum olasılık kestirimi ve Bayes kestirimi yöntemlerinden elde edilir. Nokta tahminleri elde etmek yerine, Bayesci çerçevesindeki sonsal dağılımı tahmin etmeye yönelik prosedürler nispeten yakın zamanda önerilmiştir. Böyle bir yöntem olan Gibbs örneklemesi, marjinal dağılımdan ziyade ortak sonsal dağılımı kullanarak madde ve yetenek parametrelerini kestiren bir yaklaşımdır (ör. Albert, 1992; Johnson & Albert, 1999; Kim, 2001; Patz & Junker, 1999). Madde ve yetenek parametrelerini kestirmek için kullanılabilecek Gibbs örneklemesinin birkaç farklı versiyonu ve uygulamasının olduğu unutulmamalıdır. Yine de, tüm Bayesci kestirim metotları, özellikle karşılaştırılabilir önseller kullanıldığında veya yerel olarak tekdüze önseller kullanıldığında karşılaştırılabilir madde ve yetenek parametre kestirimleri vermelidir. Bu çalışma, tek parametreli lojistik Rasch modelini kullanarak bu sorunu araştırmak için tasarlanmıştır. Özellikle, reddetme örneklemesi (GS1) kullanılan bir Gibbs örnekleme yönteminin madde güçlük ve kişi yetenek parametre kestirimleri incelenmiş ve veri artırma (GS2) yönteminin yanı sıra CML, MML ve JML kullanılan başka bir Gibbs örnekleme yöntemi ile karşılaştırılmıştır. Bu çalışmada GS1 için yeni notasyonlar kullanmak yerine Swaminathan ve Gifford'un (1982) Rasch modelinde Bayes kestirimi ile ilgili önermiş olduğu notasyon takip edilmiştir. GS1'i mevcut çalışmada farklılaştıran temel konu ve Swaminathan ve Gifford (1982)'da kullanılan uygulama, sonsal maksimizasyon ve yakınsama kavramında yatmaktadır. Bayes kestiriminde ve Gibbs örnekleme yönteminde önsellerin belirlenmesinin önemine rağmen, literatürde önsel seçimi ve kullanımı konusunda fazla bir şeffaflık olmadığı gözlenmiştir. Bu çalışma aynı zamanda, Rasch modelinde Swaminathan ve Gifford'un (1982) hiyerarşik Bayes çerçevesi bağlamında önsel seçiminin rolünü de göstermektedir.

Yöntem

Bu çalışmada Rasch modeli altında hem madde güçlük hem de kişi yetenek parametrelerinin kestirimi için bir karşılaştırma yapılmıştır. GS1, GS2, CML, MML ve JML'yi karşılaştırmak için, (1) Hukuk Fakültesi Kabul Testi 6. Bölüm (LSAT6; Bock & Aitkin, 1981; Bock & Lieberman, 1970) ve (2) Hukuk Fakültesi Kabul Testi 7. Bölüm (LSAT7) verileri kullanılmıştır. LSAT6 (1000 kişi ve 5 madde) ve LSAT7 verileri yayınlanmış birçok makale ve kitapta daha önce analiz edilmiştir (ör., Andersen, 1980; McDonald, 1999). Simülasyon verileri yerine bu verilerin kullanılması, farklı kestirim yöntemlerinin karşılaştırılmasını yapmak için okuyuculara bir temel sağlamaktadır.

Bu çalışmada GS1 kestirimleri OpenBUGS programı kullanılarak elde edilmiştir. GS2 tahminleri, Johnson ve Albert (1999)'dan gelen kodu içeren MATLAB (MathWorks, 1996) kullanılarak elde edilmiştir. LSAT6 ve LSAT7 için önce madde güçlük parametre kestirimleri daha sonra da kişi

yetenek parametre tahminleri rapor edilmiştir. CML, MML ve JML sözdizimlerini göstermek gerekli değildir. Ayrıca GS2 için Johnson ve Albert (1999, s. 248)'de sunulan MATLAB fonksiyonu herhangi bir modifikasyon olmaksızın kullanılmıştır. Bununla birlikte, OpenBUGS sözdizimini sunmak gerekli görülmüştür. Sözdiziminin gerekli bölümleri Ekte yer almaktadır. Ekte, hiperparametre varyansının tersi hem yetenek hem de madde güçlük parametreleri için dgamma (a = 2.5, b = 5) ile belirtilmiştir. Bu önsel belirleme Swaminathan ve Gifford'un (1982) $\nu = 5$ ve $\lambda = 10$ değerlerine eşdeğerdir. Ayrıca, güçlük parametrelerinin başlangıç değerleri için, p_j olarak gösterilen klasik madde güçlüğünün log oranlarının ortalanmış değerinin (yani, 0'da ortalanmış olan log $[(1-p_j) / p_j]$ değerleri) kullanıldığı dikkate alınmalıdır. Yetenek parametreleri için benzer başlangıç değerleri belirtilmiştir.

Kim ve Bolt (2007) ve Kim (2001)'in önerilerine dayanarak burn-in kısmındaki tekrar sayısı 1000'e ayarlanmış ve sonraki 10,000 tekrarı simüle edilmiş çekilişlerin yakınlaşmasını gösteren sonsal dağılımları oluşturmak için GS1 ve GS2'de kullanılmıştır (bkz. Gilks, Richardson & Spiegelhalter, 1996). Zincirlerin yakınsaklığı, geçmiş ve otokorelasyon çizimleri kontrol edilerek görsel olarak izlenmiştir. GS1 veya GS2'deki örneklenmiş değerleri özetlemenin birçok farklı yolu olduğuna dikkat edilmelidir. Gerçek sonsal güvenilirlik aralığını kullanmak yerine, bu çalışmada sonsal ortalamalar ve sonsal standart sapmalar kullanılmıştır. İlgili parametreler için örneklerin marjinal sonsal yoğunlukları, GS1'de tek modlu ve normal dağılım göstermiştir. GS2'de örneklenen değerler de benzer sonuçlar vermiştir.

Sonuç ve Tartışma

Bu çalışmada farklı kestirim metotları ve farklı önsel dağılımlar aynı veriler üzerinden karşılaştırılmıştır. LSAT6 verisi ile elde edilen madde güçlük parametresi tahminleri ve eşlik eden standart hatalar veya sonsal standart sapmalar arasında bazı farklılıklar gözlenmiştir. Bu bulgular arasında en belirgin olanı GS1 ve MML/EAP kestirimlerinin çok benzer çıkmasıdır. Diğer kestirim yöntemleri küçük test büyüklüğü nedeniyle biraz farklılık göstermiştir. LSAT7 verileri için, tüm metotlar, LSAT6 verileri için olduğu gibi, madde güçlük kestirimleri için pratik olarak aynı sonuçları vermiştir. Önsel belirlemelerin (prior specifications) madde güçlük ve yetenek parametre kestirimleri üzerindeki etkilerini değerlendirmek için, aynı LSAT6 ve LSAT7 verileri OpenBUGS ile analiz edilmiştir. LSAT6 ve LSAT7 verileri için, önsel belirlemelerin hepsi, madde güçlük tahminleri için pratik olarak aynı sonuçları vermiştir, fakat yetenek tahminleri için biraz farklı sonuçlar elde edilmiştir.

İki Gibbs örnekleme yöntemi, GS1 ve GS2, arasındaki ana fark, hem önsel dağılımların özelliklerinde hem de temel örnekleme prosedürlerinde yatmaktadır. GS1'de kullanılan önsel dağılımlar, Swaminathan ve Gifford (1982)'un önerisini takip eden hiyerarşik forma sahiptir. Örneğin, yetenek parametresinin normal olan önsel dağılımına ait hiperparametrenin ortalaması, bilgi-verici olmayan (non-informative) bir tekdüze dağılıma sahip iken önsel normal olanın hiperparametre varyansının tersi, bir gama dağılımına sahiptir. Gama (a = 2.5, b = 5) dağılımlı GS1'de, gama dağılımının önsel örneklem büyüklüğü 2*(2.5) = 5 olarak belirlendi ve önsel beklenen değer 2.5 / 5 = 0.5 idi (yani, hipermetre varyansının beklenen değeri 5 / 1.5 = 3.33). Bu önsel belirlemenin, Swaminathan ve Gifford'un (1982) $\nu = 5$ ve $\lambda = 10$ değerlerine eşdeğer olduğunu unutmayın. Swaminathan ve Gifford ikinci bir çalışmada, eşdeğer belirlemelere dönüştürülmüş olan başka üç özellik daha kullanmıştır. Bu çalışmada Gamma (2.5, 5) kullanımı makul bir seçenek olarak görünmektedir. Swaminathan ve Gifford (1982) da benzer sonuçları raporlamıştır. Hiyerarşik formda önselleri kullanmanın yanında Rasch modeli için önselleri belirlemenin başka yolları da vardır (bkz. Kim, 2001; Levy & Mislevy, 2016; Spiegelhalter ve ark., 1996b; Stone & Zhu, 2015). Johnson ve Albert'ın (1999) item_r1 fonksiyonunda GS2 için, önsel teta hiperparamatreleri standart bir normal dağılıma ayarlanmış, öte yandan standart sapma parametrelerinin birliği olarak ayarlanmıştır. GS2'ye ait ayrıntılı Gibbs örneklemesi için Johnson ve Albert (1999, s. 202–204)'e bakılabilir. Dolayısıyla GS1 ve GS2 sadece modelin matematiksel formlarında değil, aynı zamanda kullanılan önsellerde de farklılık göstermektedir.

Appendix: OpenBUGS Code

```
model {
# patterned data to individual responses
  for (i in 1:cof[1]) {
    for (j in 1:J) { x[i, j] <- pattern[1, j] }
  }
  for (g in 2:G) {
    for (i in cof[g-1]+1:cof[g]) {
      for (j in 1:J) { x[i, j] <- pattern[g, j] }
    }
  }
# Rasch model
  for (i in 1:I) {
    for (j in 1:J) {
      logit(p[i, j]) <- theta[i] - beta[j]
      x[i, j] ~ dbern(p[i, j])
    }
# ability prior
    theta[i] ~ dnorm(mut, taut)
    t[i] <- theta[i] - mean(beta[])
  }
# item prior
  for (j in 1:J) {
    beta[j] ~ dnorm(mub, taub)
    b[j] <- beta[j] - mean(beta[])
  }
# hyperpriors
  mut ~ dunif(-5, 5)
  taut ~ dgamma(2.5, 5)
  phit <- 1 / sqrt(taut)
  mub ~ dunif(-5, 5)
  taub ~ dgamma(2.5, 5)
}

# lsat6 patterned data with cumulative observed frequencies
list(I = 1000, G = 32, J = 5,
  cof = c(3, 9, 11, 22, 23, 24, 27, 31, 32, 40,
    40, 56, 56, 59, 61, 76, 86, 115, 129, 210,
    213, 241, 256, 336, 352, 408, 429, 602, 613, 674,
    702, 1000),
  pattern = structure(.Data = c(
    0, 0, 0, 0, 0,
    0, 0, 0, 0, 1,
    0, 0, 0, 1, 0,
    0, 0, 0, 1, 1,
    0, 0, 1, 0, 0,
    0, 0, 1, 0, 1,
    0, 0, 1, 1, 0,
    0, 0, 1, 1, 1,
    0, 1, 0, 0, 0,
    0, 1, 0, 0, 1,
    0, 1, 0, 1, 0,
    0, 1, 0, 1, 1,
    0, 1, 1, 0, 0,
    0, 1, 1, 0, 1,
    0, 1, 1, 1, 0,
    0, 1, 1, 1, 1,
    1, 0, 0, 0, 0,
    1, 0, 0, 0, 1,
    1, 0, 0, 1, 0,
```

```
    1, 0, 0, 1, 1,
    1, 0, 1, 0, 0,
    1, 0, 1, 0, 1,
    1, 0, 1, 1, 0,
    1, 0, 1, 1, 1,
    1, 1, 0, 0, 0,
    1, 1, 0, 0, 1,
    1, 1, 0, 1, 0,
    1, 1, 0, 1, 1,
    1, 1, 1, 0, 0,
    1, 1, 1, 0, 1,
    1, 1, 1, 1, 0,
    1, 1, 1, 1, 1), .Dim = c(32, 5))
)

# initial values
list(
  beta = c(-1.163685322, 0.44376115, 1.121494003, 0.165095519, -0.566665352),
  mut = 0, taut = 1,
  mub = 0, taub = 1,
  theta = c(-2.1972246, -2.1972246, -2.1972246, -1.3862944, -1.3862944,
  .
  .
  .
  2.1972246) # 1000 initial theta values
)
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3); 277-294

The Examination of Item Difficulty Distribution, Test Length and Sample Size in Different Ability Distribution

Melek Gülşah ŞAHİN* Yıldız YILDIRIM **

Abstract
This is a post-hoc simulation study which investigates the effect of different item difficulty distributions, sample sizes, and test lengths on measurement precision while estimating the examinee parameters in right and left-skewed distributions. First of all, the examinee parameters were obtained from 20-item real test results for the right-skewed and left-skewed sample groups of 500, 1000, 2500, 5000, and 10000. In the second phase of the study, four different tests were formed according to the *b* parameter values: normal, uniform, left skewed and right skewed distributions. A total of 80 conditions were formed within the scope of this research by selecting 20-item and 30-item condition as the test length variable. In determining the measurement precision, the RMSE and AAD values were calculated. The results were evaluated in terms of the item difficulty distributions, sample sizes, and test lengths. As a result, in right-skewed examinee distribution, the highest measurement precision was obtained at the normal *b* distribution and the lowest measurement precision was obtained at the right skewed *b* distribution. A higher measurement precision was obtained in the 30-item test, however, it was observed that the change in the sample size didn't affect the measurement precision significantly in right-skewed examinee distribution. In the left skewed distribution, the highest measurement precision was obtained at the normal *b* distribution and the lowest measurement precision was obtained at the left-skewed *b* distribution. Also it was observed that the change in the sample size and test length didn't affect the measurement precision significantly in the left-skewed distribution.

Key Words: Item response theory, examinee distribution, item difficulty distribution, sample size, test length.

INTRODUCTION

During the phases of development and scoring process of the tests used to recognize individuals in the fields of Education and Psychology, Classical Test Theory (CTT) and Item Response Theory (IRT) are utilized. These two theories are considered fundamentals in the field of measurement and evaluation. While IRT emerged through the midst of 20th century, the history of CTT dates back to the earlier ages (Crocker & Algina, 1986). IRT is an advantageous and powerful approach in test development, item analysis, and scoring processes (Thompson & Weiss, 2011). Unlike CTT, it is considered that there is a relation between the responses given and the characteristics that the test measures in IRT, and this relation is shown with an increasing function that is named as Item Characteristic Curve (ICC). As IRT does not vary from one group to another, the parameters that determine this curve will remain the same (Lord & Novick, 1968). There are four parameters in the definition of IRT. These are item discrimination parameter (a), item difficulty parameter (b), pseudo guessing parameter (c), and upper asymptote (d). Also, the mathematical equations that describe ICC form IRT models. In addition, the performance of each person who responses the items in the test can be estimated through the instrumentality of the factors named such as characteristics, latent trait or ability (Hambleton, Swaminathan & Rogers, 1991). Another term in the theory is item information function and test information function. The contribution of any item in the scale to the accuracy of measurement done with the whole scale is determined through item information function. Moreover, the test information function is obtained through the total amount of item information function.

*Instructor Dr., Gazi University, Gazi Education Faculty, Ankara-Turkey, e-mail: melekgulsah@gmail.com, ORCID ID: https://orcid.org/0000-0001-5139-9777
** Research Assistant, Adnan Menderes University, Education Faculty, Aydın-Turkey, e-mail: yildiz.yildirim@adu.edu.tr, ORCID ID: https://orcid.org/0000-0001-8434-5062

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, Cilt 9, Sayı 3, Sonbahar 2018, 277-294.
Journal of Measurement and Evaluation in Education and Psychology, Vol. 9, Issue 3, Autumn 2018, 277-294.
DOI: 10.21031/epod.385000
Received: 28.01.2018
Accepted: 06.08.2018

Item information function and test information function can be obtained independently of sample of individuals. Moreover, these functions are related to standard error of measurement at any ability levels. Due to this features of item information function and test information function is considered as an alternative to reliability and standard error in CTT. The average of test information function at all ability levels means the "reliability" coefficient (marginal reliability) (Hambleton & Swainathan, 1985).

Unidimensionality, local independence and normality assumptions are found in the unidimension and parametric models of IRT. Unidimensionality assumption is based on the statistical independence among items (Crocker & Algina, 1986) and test items measure only one ability (Hambleton et al., 1991). Local independence assumption is related to unidimensionality and it means that, when the abilities influencing the test performance of the individuals are at the same level, individuals' responses to any pair of items are statistically independent from the responses to any other test items. Although unidimensionality and local independence are different terms, when the test ensures its unidimensionality, it means that the local independence assumption is obtained (Hambleton et al., 1991).

The characteristic features of IRT has improved test development, test bias identification, test equating and the limitations have been removed in these conditions (Hambleton & Swaminathan, 1985). Thanks to the advantages of IRT, this theory has been preferred in the examinations especially like PISA (The OECD Programme for International Student Assessment) and TIMSS (The Trends in International Mathematics and Science Study) (Martin, Mulis & Hooper, 2016; OECD, 2017). In addition, it is seen in many national and international research that test results are evaluated within the context of IRT (Ackermann, 1994; Bhakta, Thennant, Horton, Lawton & Andrich, 2005; Çelen & Aybek, 2013; İlhan, 2016). The exams used in education are prepared for many different purposes, and these exams are extremely important for individuals. These purposes can include student selection and placement, proficiency, diagnostic tests etc. These tests will have various psychometric characteristics depending on the purpose of development, the characteristics of individuals or the number of individuals taking the test. For example, if the number of students are more but the number of the students to be selected according to the results is less, the test can be expected to be difficult. However, if the test is to be developed to diagnose the existing knowledge (not to select and place), the test is expected to be easier than selection and placement tests and to consist of items with moderate difficulty, if possible. It is more important here to identify how the validity and reliability will be affected in the tests that have different item difficulty index. In addition, how the ability distribution of the individuals that take the test affect the validity and reliability should also be identified. In this study, based on the results of a national exam, the effect of test length and sample size for different ability distributions in the tests that have different *b* parameters within ability parameter estimation on measurement precision was analyzed.

In the literature, there are studies that analyze the effect of sample size on measurement precision in various models and items with different scores in the item response theory (Boughton, Klinger & Gierl, 2001; Cheng & Yuan, 2010; De Ayala & Bolesta, 1999; DeMars, 2002; DeMars, 2003; Montgomery & Skorupski, 2012; Preston & Reise, 2014). In addition to these, there are studies which consist at least two of sample size, test length and ability distribution type conditions. (Ankenmann ve Stone, 1992; Baker, 1998; Guyer ve Thompson, 2011; Hulin, Lissak ve Drasgow, 1982; Kieftenbeld ve Natesan, 2012; Lautenschlager, Meade ve Kim, 2006; Preinerstorfer ve Formann, 2012; Roberts ve Laughlin, 1996; Seong, Kim ve Cohen, 1997; Stone, 1992; Swaminathan ve Gifford; 1979; Wang ve Cheng, 2005; Wollack, Bolt, Cohen ve Lee, 2002). Furthermore, while there are studies that *a* parameter is obtained within different ranges and that analyze its impact on measurement precision (DeMars, 2003; Preston & Reise, 2014; Reise & Yu, 1990), fewer studies examine *b* parameters' impact on measurement precision. Some studies related to this study are summarized as follows.

Lautenschlager et al. (2006), in a post-hoc simulation study within graded response model (GRM), examined the effect of 7 different sample sizes (75, 150, 200, 300, 500, 1000 and 2000 individual), four different test lengths (5, 10, 15 and 20 items), and three different sample distributions (normal,

skewed and uniform) on ability and item parameter estimation. The researchers used maximum posteriori (MAP) estimation method in the ability parameter estimation. In the study, the results showed that sample size does not change the root mean squared error (RMSE) values but RMSE values decreased when the test length increases. Ankenmann and Stone (1992) carried out a post-hoc simulation study using three different test lengths (5, 10, and 20 items), with a sample size of 125, 150, 500 for one-parameter GRM and with a sample size of 250, 500, and 1000 for 2-parameter GRM, they analyzed how ability estimation was affected. The researchers that used marginal maximum likelihood (MML) in parameter estimation used MULTILOG Program. As a result, it was concluded that sample size did not have an important effect on ability parameter estimation. In addition, it was found that the longer the test length is, the more precise the measurement in ability estimation. Kieftenbeld and Natesan (2012) conducted another post-hoc simulation in their study using a four different test lengths (5, 10, 15, and 20 items), five different sample sizes (75, 150, 300, 500, and 1000 individuals) and three different ability distribution types (normal, uniform, and skewed), and they analyzed the effect of these conditions on ability and item parameter. In the study, MML and Markov Chain Monte Carlo (MCMC) methods were used for estimation. They conducted the study within the context of GRM and estimated the parameters using MULTILOG program. The results of the study revealed that test length described the highest variance in RMSE whereas sample size described a less amount of the variance. Preinerstorfer and Formann (2012) analyzed the effect of two different sub-groups (1 and 2 sub-group), homogeneity and heterogeneity of the groups, four different test lengths (10, 15, 25 and 40 items) and three sample sizes (500, 1000, and 2500) on measurement precision in parameter estimation using mixed Rasch model. As a result, it was found that as sample size and test length increased, so did the measurement precision.

In the literature, for the models related to polytomous items and Rasch model, there are some studies that analyze the effect of sample size and/or test length on measurement precision, and some other similar studies with logistic models related to dichotomous items. For example, Swaminathan and Gifford (1979) analyzed the effect of ability and item parameter estimation on measurement precision using Urry and MLE methods. They used different test lengths (10, 15, 20, and 80), different sample sizes (50, 200, and 1000), and different ability distribution types (normal, uniform, and skewed) within 3PL model. As a result, they stated that when the sample size and test length increased, so did the measurement precision within ability parameter, and there was a little effect of sample size on measurement precision. Hulin et al. (1982) carried out a Monte-Carlo study using 2PL and 3PL models and analyzed the effect of different sample sizes (200, 500, and 1000), different test lengths (15, 30, and 60) on measurement precision within item and ability parameter estimation. The result of the study revealed that the accuracy of ability estimation in 3PL is less in small samples and small lengths. In addition, it was found that the sample size in 30 and 60 item tests in 3PL model did not affect RMSE and correlation values much. Stone (1992) analyzed the effect of different sample sizes (250, 500, and 1000), different test lengths (10, 20, and 40) and different distribution types (normal, skewed, and platykurtic) in 2PL model on measurement precision within parameter estimation. The result of the study revealed that the most significant condition that affected measurement precision was test length within ability parameter estimation (especially among extreme ability parameters). In addition, it was found that when the test length gets longer, error of estimation decreased significantly. Furthermore, they also found that the increase in the sample size did not reduce the deviation. Stone also analyzed the measurement precision within item level and the effect of research conditions when *b* parameter was in different levels (average (0, 02), easy (-2, 18), difficult (1, 82)) on measurement precision. In this context, it was found that when the item difficulty was average, lower RMSE values were achieved within item parameter estimation, and the highest RMSE values were seen in easy items. Cheng and Yuan (2010) aimed to correct the standard error of ability estimation using MLE method within 2PL model. These researchers, who analyzed the effect of sample size on standard error, determined the sample size as 200 and 2000. It was found that the increase in the sample size did not affect the standard error significantly.

Finally, some studies that analyze the effect of sample size and test length on measurement precision are summarized below. Köse (2010) aimed to analyze the effect of different sample sizes (500, 1000,

and 1500) and different test lengths (12 and 24) on item and ability parameter estimation and model data fit in unidimensional (2PL) and multidimensional models. The results of the study reveal that sample size in ability parameter estimation did not have a significant effect on both unidimensional and multidimensional models. In addition, Köse stated that, based on RMSD values, the increase in the number of items in ability parameter estimation caused less defective results. Koğar (2015) carried out a Monte Carlo study using unidimensional, unidimensional non-parametric and multi-dimensional IRT models and analyzed the effect of different sample sizes (100, 500, 1000, and 5000), different test lengths (5, 15, and 25) and different inter-dimensional correlation values (0,00, 0,25, and 0,50) on item parameter estimation and model fit. The results suggested that, in unidimensional and multidimensional models, in order for the item parameter estimation to be more accurate, the sample size and test length should be greater.

In the literature, the studies usually focus on analyzing the effect of some variables such as sample size, test length, and item discrimination index on measurement precision within ability parameter estimation. Different from many studies, this study investigated how the measurement precision of the ability parameter estimation is affected by different *b* parameter distributions (normal, uniform, right-skewed, and left-skewed), in addition to analyzing the effect of sample size and test length in left and right skewed ability distributions.

Purpose of the Study

This study aims to analyze the effect of different *b* parameter distributions, test lengths, sample sizes on measurement precision of ability parameter estimation in right skewed and left-skewed ability distributions. It was found that literature generally focuses on different conditions that affect measurement precision within ability parameter estimation. As stated in the introduction part of this study, the studies usually analyze the effect of sample size and test length on measurement precision. However, no studies were found in literature that analyze the effect of different *b* parameter distributions on measurement precision in the groups that have different ability distributions, different test lengths and sample sizes. Production of four different tests based on different item difficulty distributions is considered important. The problem of the study is "what is the effect of different item difficulty distributions, sample sizes, and test lengths in right-skewed and left-skewed ability distributions on measurement precision of ability parameter estimation?"

Sub-problems of the study are as follows:

1. What is the effect of different test lengths, sample sizes, item difficulty distributions within right-skewed ability distribution on measurement precision of ability parameter estimation?

2. What is the effect of different test lengths, sample sizes, item difficulty distributions within left-skewed ability distribution on measurement precision of ability parameter estimation?

METHOD

Data Production

Obtaining Ability Parameter Values

In this post-hoc simulation study, real data were used to collect ability parameters. The real data were obtained from the 20-items mathematics subtest of Placement Test (Seviye Belirleme Sınavı-SBS) applied in 2012. This placement test was used to select students who will continue high school education. In the study, totally five sample sizes (500, 1000, 2500, 5000, and 10000) were chosen from the data set. Previous studies in the literature (Ankenmann & Stone, 1992; Baker, 1998; DeMars, 2002; Guyer & Thompson, 2011; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Montgomery & Skourpski, 2012; Preinerstorfer & Formann, 2012; Preston & Reise; 2014; Reise & Yu, 1990; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979; Thissen & Wainer, 1982; Wang & Cheng, 2005; Wollack et

al., 2002, Yavuz & Hambleton, 2016) were utilized while choosing the sample size. For each sample size chosen for obtaining the ability parameters, both right-skewed and left-skewed ability, distributions were chosen from the real data. During the selection of right and left-skewed distributions for each sample size for the right-skewed distribution, SBS data, which is originally a right-skewed data set (coefficient of skewness=1,05), was done randomly. For the left-skewed data sets, similar to the study of Doğan and Tezbaşaran (2003), intended sample distribution was achieved through purposive sampling, and the groups whose coefficient of skewness is ≈-1,00 were chosen for all sample sizes.

Similar to the coefficient of skewness values used in Doğan and Tezbaşaran (2003), Bahry (2012) and Sen (2014), it was determined the coefficient of skewness as +1,00 in this study. For the left-skewed distribution, Doğan & Tezbaşaran (2003) and Bıkmaz Bilgen & Doğan (2017) used a -1,00 coefficient of skewness in their studies. After these groups were chosen from the areal data, maximum likelihood estimation method was used in MULTILOG 7.03 program (Thissen, Chen & Bock, 2003) and the groups' ability parameters were estimated with 25 replications, and this post-hoc simulation study was completed.

Simulation of Item Parameters

In the second step of the study, different four tests were created which have different *b* parameters: tests with normal distribution, uniform distribution, right-skewed and left-skewed distribution. The statistics used in test development were determined according to the values and suggestions within the studies in the literature (Ankenmann & Stone, 1992; Baker, 1998; Bahry, 2012; De Ayala & Sava-Bolesta,1999; DeMars, 2002; DeMars, 2003; Dolma, 2009; Fotaris, Mastoras, Mavridis & Manitsaris, 2010; Han, 2012; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Montgomery & Skourpski, 2012; Preston & Reise; 2014; Reise & Yu, 1990; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979). In accordance with these studies, *a* parameter value was determined as min=0,5 and max=2 in the simulation of item parameters, and *c* parameter value was determined as min=0 and max=0,05. Four different item difficulty distribution were created for left-skewed *b* parameter α=8; β=2; for right-skewed *b* parameter distribution α=2; β=8; for uniform *b* parameter distribution min=-3; max=+3 and for normal *b* parameter distribution average=0 and sd=1 values were used. For the test length variable of the study, two different conditions with 20 and 30 items were determined. The reason why the test length was determined as 20 and 30 items is that these test lengths are mainly used in national exams and the studies in the literature use similar test lengths (Ankenmann & Stone, 1992; Baker, 1998; Boughton et al., 2001; Craig & Kaiser, 2003; DeMars, 2003; Fotaris et al., 2010; Guyer & Thompson, 2011; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979; Wang & Cheng, 2005; Wollack et al., 2002, Yavuz & Hambleton, 2016). 80 conditions (2 ability distribution, x5 sample size, x4 *b* parameter distribution, x2 test length) dealed within the scope of the study were created via WinGen 3 program (Han, 2007; Han & Hambleton, 2007) after 25 replications. Within the scope of the study, the reason why 25 replications were made is that it is a sufficient number in the elimination of sample bias (Harwell, Stone, Hsu & Kirisci, 1996).

Data Analysis

During data analysis process, firstly ability parameter estimation produced data were done through MULTILOG 7.03 and 2000 times (80 conditions x 25 replication) based on MLE method. Then the estimated measurement precision of ability parameter was analyzed as parameter recovery studies in IRT generally use measurement precision calculation. To analyze measurement precision, RMSE and "average absolute deviation (AAD)" values were calculated. RMSE and AAD values were calculated after each replication and compared to the number of replications, then the average score was reported and discussed. To calculate these values, the following formulas were used:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta_j} - \theta_{Tj})^2}{N}}$$

$$AAD = \frac{\sum_{i=1}^{N}\left|\widehat{\theta_J} - \theta_{T_j}\right|}{N}$$

In these formulas, θ_{T_j} j. means actual ability parameter for the individual; $\widehat{\theta_J}$ j refers to ability parameter estimated for the individual and N describes the sample size. When RMSE and AAD values get closer to 0, the measurement precision increases. Thus, the accuracy of parameter estimation also increases. In addition, some interpretations were made according to the criterion that RMSE value is less than 0,10 (DeMars, 2003; Sen, Cohen & Kim, 2015; Tate, 2000).

RESULTS

This part represents the findings within the context of sub-problems of the study.

1. Sub-problem: What is the effect of different test lengths, sample sizes, item difficulty distributions within right-skewed ability distribution on measurement precision of ability parameter estimation?

All the RMSE and AAD values from analysis done for right-skewed ability distribution are shown in Table 1.

Table 1. RMSE and AAD Values in Right-Skewed Ability Distribution in Relation to Test Conditions

Right-Skewed Ability Distribution		Item Difficulty Parameter Distribution							
		Normal		Uniform		Left-Skewed		Right-Skewed	
Test Lengths	Sample Sizes	RMSE	AAD	RMSE	AAD	RMSE	AAD	RMSE	AAD
20	500	0,080	0,317	0,112	0,460	0,144	0,562	0,235	1,108
	1000	0,080	0,320	0,115	0,469	0,150	0,587	0,232	1,087
	2500	0,079	0,315	0,112	0,459	0,149	0,583	0,231	1,089
	5000	0,079	0,314	0,112	0,458	0,148	0,581	0,232	1,091
	10000	0,079	0,315	0,112	0,460	0,148	0,580	0,232	1,090
30	500	0,070	0,275	0,101	0,411	0,156	0,637	0,231	1,101
	1000	0,071	0,282	0,102	0,419	0,163	0,665	0,228	1,078
	2500	0,070	0,279	0,100	0,408	0,161	0,663	0,228	1,081
	5000	0,070	0,278	0,100	0,408	0,161	0,663	0,228	1,082
	10000	0,070	0,280	0,100	0,411	0,161	0,661	0,228	1.082

In Table 1, RMSE and AAD values, which were used to determine the measurement precision for 40 conditions within right-skewed distribution, are represented. In this sub-problem, the variation of RMSE and AAD values (in different *b* parameter distributions and sample size for 20 and 30 test items within the context of right-skewed ability distribution) is shown in Figure 1 and the figures are discussed with Table 1.

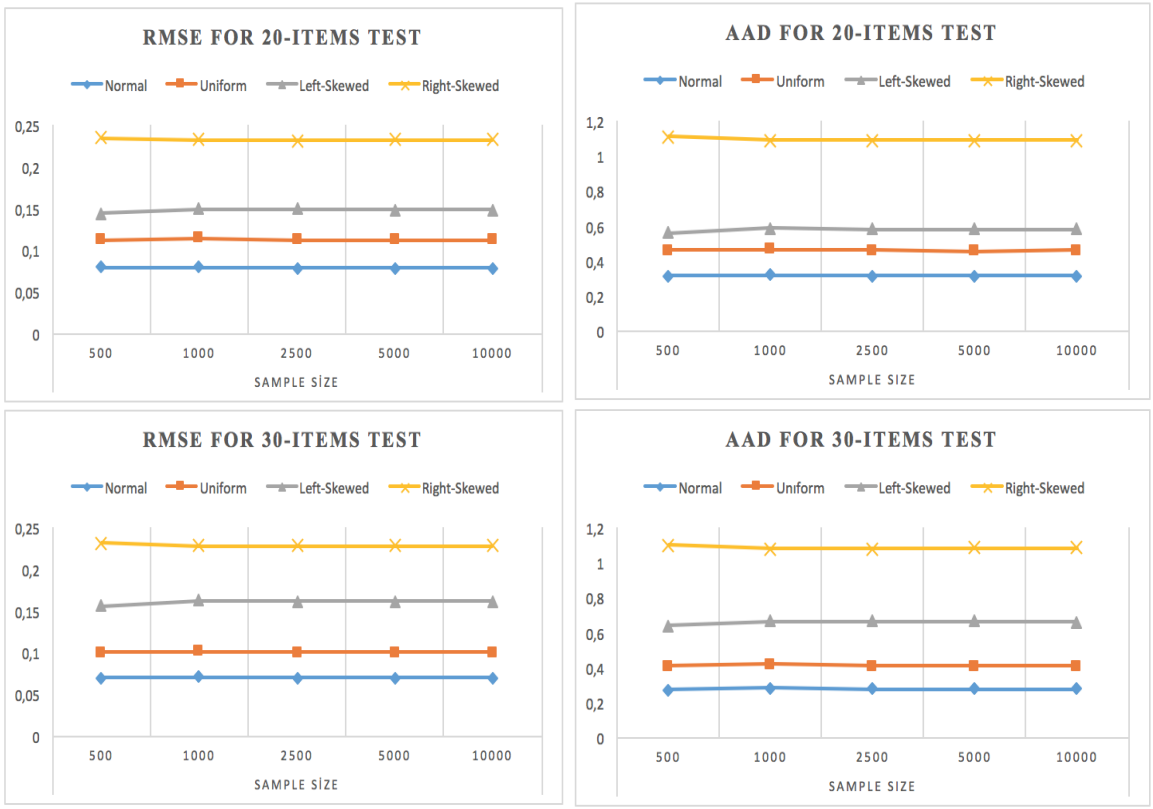


Figure 1. Graphics in Relation to RMSE and AAD within the Context of Test Length for Right-Skewed Ability Distribution.

When Figure 1 and Table 1 are analyzed, within all sample sizes (500, 1000, 2500, 5000, and 10000) that has right-skewed ability distribution, when *b* parameter distribution is normal, it can be seen that the lowest RMSE and AAD values were obtained for both 20-item test and 30-item test. These RMSE and AAD values are followed by uniform and left-skewed distribution for all sample sizes respectively. However, the highest RMSE and AAD values were obtained from the distribution in which *b* parameter has right-skewed distribution. Based on these values of RMSE and AAD statistics, it can be stated that, within all sample sizes, the measurement precision is the highest when *b* parameter has a normal distribution and the lowest when it has right-skewed distribution, and the second highest measurement precision distribution type is the uniform distribution. In addition, sample size did not have much effect on RMSE and AAD values within ability parameter estimation within different *b* parameter distribution and test lengths for right-skewed ability parameter. This result can be seen in Figure 1 and Table 1. In other words, sample size did not have a significant effect on measurement precision within ability parameter estimation.

With reference to the values in Table 1, the variation of RMSE and AAD values within different *b* parameter distributions and test lengths (individually for each sample size) is shown in Figure 2 and the figures are discussed with Table 1.

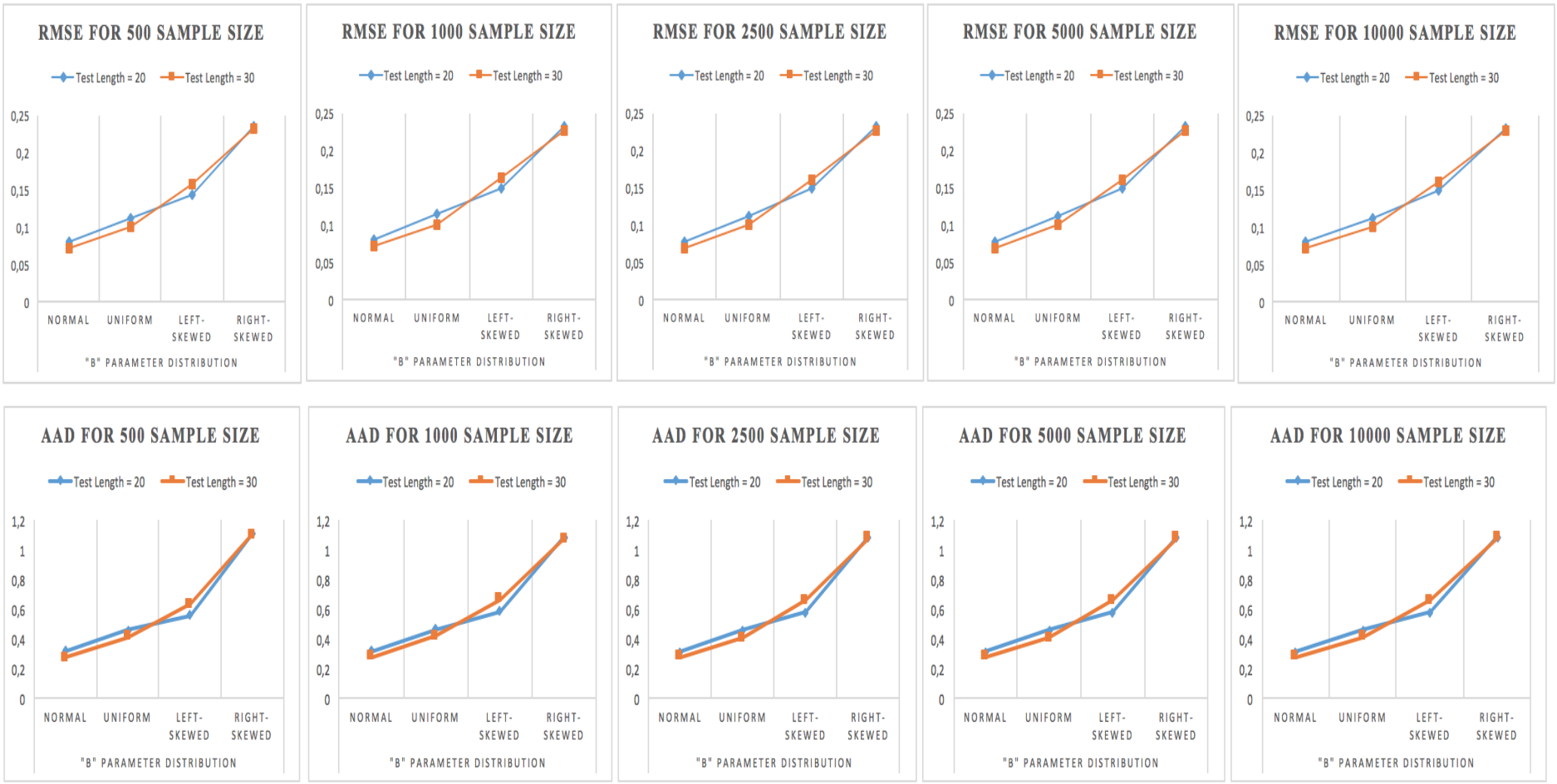


Figure 2. Graphics in Relation to RMSE and AAD Values within the Context of Sample Size for Right-Skewed Ability Distribution.

When Figure 2 and Table 1 is examined, when b distribution is normal, it can be seen that the lowest RMSE and AAD values were obtained in 30-items test. Higher RMSE and AAD values were obtained for 20 items within each sample size than the values within 30-item test. When item difficulty parameter has uniform and right-skewed distribution, for all sample sizes, the lowest RMSE and AAD values, similar to the distribution in normal item difficulty, was seen within 30-item test. Accordingly, it can be said that, in the tests that have normal, uniform, and right-skewed *b* parameter distribution, for all sample sizes, when the test length increases, the measurement precision also increases. However, for the left-skewed *b* parameter distribution, when all sample sizes are considered, the lowest RMSE and AAD values were obtained from 20-item test. It was different from the other item difficulty distributions. This may be because of the increase in the number of items with high item difficulty. Overall, when the test length increases, RMSE and AAD values decrease; and hereby measurement precision increases. When the values for right-skewed ability parameter are analyzed, it was found that, for all *b* parameter distributions, the values obtained from different test lengths were more or less the same. However, it was also seen that, in contrast with sample size, the values varied when test length changes. In conclusion, it can be stated that, based RMSE<0,10 on the criteria that Tate (2000), DeMars (2003) and Sen et al. (2015) used, all test lengths and sample sizes were convenient when the *b* parameter distribution is normal. However, in other *b* parameter distributions, all of test lengths and sample sizes were not found appropriate based on the criterion.

2. Sub-problem: What is the effect of different test lengths, sample sizes, item difficulty distributions within left-skewed ability distribution on measurement precision of ability parameter estimation?

All RMSE and AAD values obtained from the whole analysis for left-skewed ability distribution are shown in Table 2.

Table 2. RMSE and AAD Values in Left-Skewed Ability Distribution in Relation to Test Conditions

Left-Skewed Ability Distribution		Item Difficulty Parameter Distribution							
		Normal		Uniform		Left-Skewed		Right-Skewed	
Test Length	Sample Size	RMSE	AAD	RMSE	AAD	RMSE	AAD	RMSE	AAD
20	500	0,079	0,324	0,137	0,610	0,246	1,166	0,149	0,652
	1000	0,079	0,326	0,136	0,610	0,248	1,183	0,147	0,656
	2500	0,079	0,326	0,138	0,616	0,250	1,191	0,146	0,638
	5000	0,079	0,328	0,137	0,611	0,250	1,192	0,146	0,640
	10000	0,079	0,327	0,138	0,617	0,250	1,191	0,146	0,639
30	500	0,078	0,322	0,137	0,610	0,248	1,176	0,150	0,656
	1000	0,079	0,327	0,137	0,615	0,249	1,184	0,147	0,643
	2500	0,079	0,327	0,135	0,604	0,250	1,191	0,146	0,641
	5000	0,079	0,327	0,138	0,617	0,249	1,189	0,146	0,639
	10000	0,079	0,326	0,138	0,617	0,250	1,190	0,146	0,639

In Table 2, RMSE and AAD values, which were used to determine the measurement precision for 40 conditions within left-skewed distribution, are represented. In the second sub-problem, the variation of RMSE and AAD values (in different b parameter distributions and sample size for 20 and 30 test items within the context of left-skewed ability distribution) is shown in Figure 3 and the figures are discussed with Table 2.

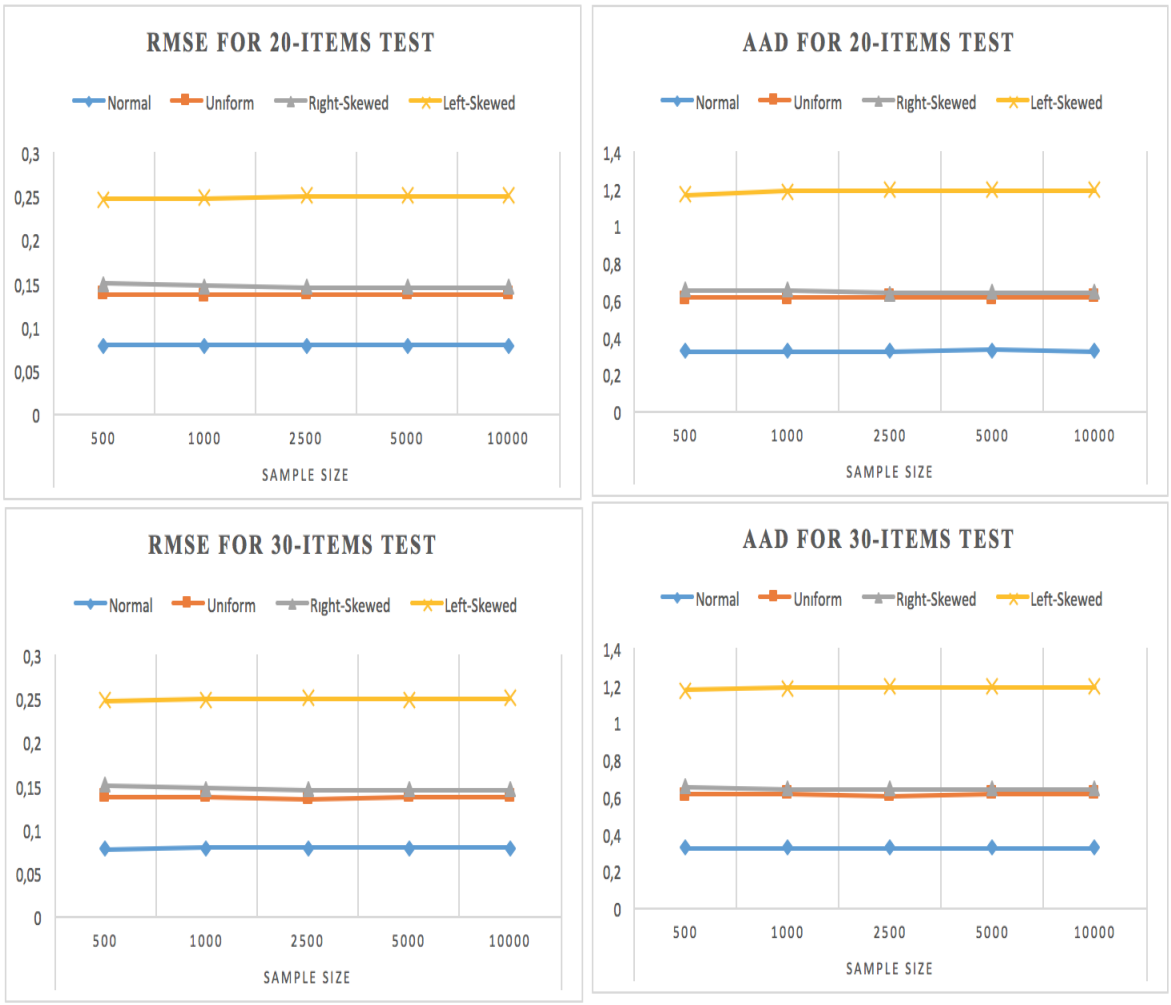


Figure 3. Graphics in Relation to RMSE and AAD Values within the Context of Test Length for Left-Skewed Ability Distribution.

When Figure 3 and Table 2 is examined, when b distribution is normal, within all sample sizes that have left-skewed ability distribution, it can be seen that the lowest RMSE and AAD values were obtained for both 20-items test and 30-item tests. These values are followed by uniform b distribution and right-skewed distribution respectively. The highest RMSE and AAD values were obtained from the distribution in which *b* parameter has left-skewed distribution. Based on these values of RMSE and AAD statistics, it can be stated that, within all sample sizes, the measurement precision is the highest when *b* parameter has a normal distribution and the lowest when it has left-skewed distribution, and the second highest measurement precision distribution type is the uniform distribution. In addition, sample size did not have much effect on RMSE and AAD values within ability parameter estimation within different *b* parameter distribution and test lengths for left-skewed ability parameter distribution. This result can be seen in Figure 3 and Table 2. In other words, sample size did not have a significant effect on measurement precision within ability parameter estimation. The variation of RMSE and AAD values within different *b* parameter distributions and test lengths (individually for each sample size) is shown in Figure 4 and the figures are discussed with Table 2.

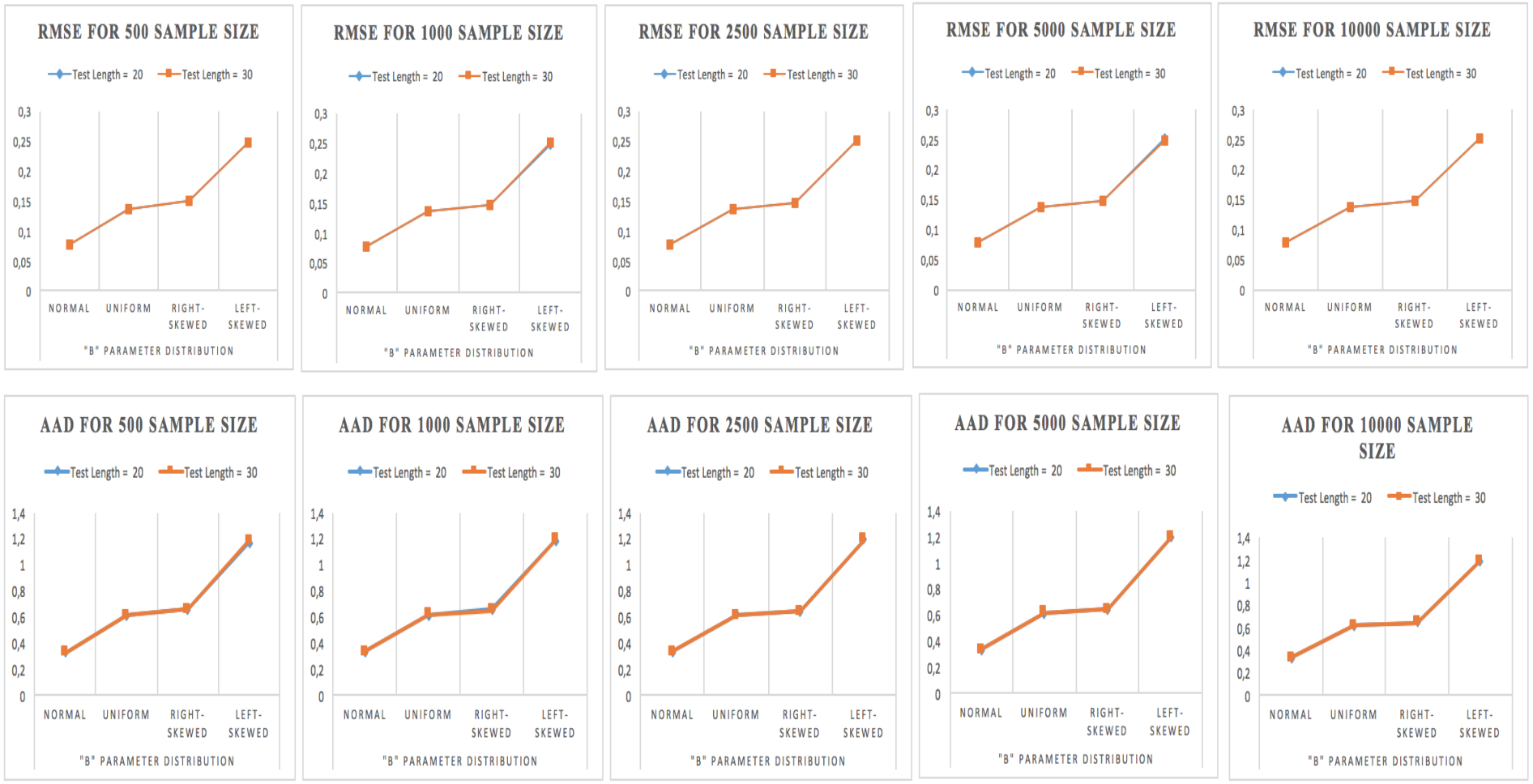


Figure 4. Graphics in Relation to RMSE and AAD Values within the Context of Sample Size for Left-Skewed Ability Distributio

When Figure 4 and Table 2 is analyzed, for left-skewed ability parameter distribution, it was seen that RMSE and AAD values are similar in both 20-item and 30 item within all item difficulty parameter distributions and sample sizes. Accordingly, it can be said that, within all sample sizes and item difficulty parameter distributions, measurement precision does not change significantly although the test length increases. In conclusion, it can be stated that, based RMSE<0,10 on the criteria that Tate (2000), DeMars (2003) and Sen et al. (2015) used, all test lengths and sample sizes were convenient when the *b* parameter distribution is normal. However, in other *b* parameter distributions, all of test lengths and sample sizes were not found appropriate based on the criterion.

DISCUSSION and CONCLUSION

In this study, measurement precision of ability parameter estimation obtained from the conditions that are generated from two different ability distribution, five different sample size, four different *b* parameter distribution, and two different test length is analyzed. The ability parameter values were estimated according to the conditions addressed by the data from a national exam. To determine the test lengths, the average test lengths of national exams were considered. To create the tests, it is considered that the conditions in which *b* parameter comprised of normal, uniform, right-skewed, and left-skewed distributions.

When the results for right-skewed ability distribution are examined, it is seen that, when the sample size of each test that has different *b* parameter distribution increases, RMSE and AAD values that are measured for measurement precision do not change significantly. When the effect of sample size change for 20-items and 30-items tests is examined, it is seen that RMSE and AAD values decrease when sample size increases. However, when the conditions in which sample size and test length has different *b* parameter distributions, the best results were obtained when *b* parameter has normal distributions. This condition is followed by the condition which *b* parameter has uniform distribution. In the conditions that has uniform distribution, similar to other conditions, there is not a significant effect of different sample sizes on measurement precision. When *b* parameter had left-skewed distribution, RMSE and AAD values did not vary much in different sample sizes but they decreased when test length increased. Lower RMSE and AAD values were obtained for 30 items than 20-items test when *b* parameter distribution had right-skewed. In addition, it can be stated that, when sample size increases, RMSE and AAD values do not vary significantly but the difference between 500 and 1000 individuals are higher than other sample sizes. In right-skewed b distribution, RMSE and AAD values were higher than other b distributions. Similarly, Stone (1992) compared normal ability distribution for easy items and right-skewed ability distribution and found that right-skewed ability distribution (such conditions as 20 items and 500-1000 sample size) had lower measurement precision values than normal ability distribution.

When left-skewed ability distribution was examined, it is seen that, when sample size for each test that has different *b* parameter increased, RMSE and AAD values did not have significant change. When the effect of test length was analyzed, it was found that in the group that had left-skewed ability parameter, the increase of the test length did not affect measurement precision in general. When the effect of item difficulty parameter was examined, it was found that the lowest RMSE and AAD values were obtained when *b* parameter had normal distribution. This distribution was followed by uniform *b* parameter distribution (relevant for both test lengths and all sample sizes). It was found that by achieving the highest RMSE and AAD values in left-skewed *b* parameter distribution and measurement precision was the lowest for these values.

The overall results of the study showed that, within both left-skewed and right-skewed ability parameter distribution, when the sample size within each *b* parameter distribution types increases, no significant change was observed in measurement precision. In the literature, some studies show the same results for similar conditions. Hulin et al. (1982) and Swaminathan and Gifford (1979), for example, stated that sample size does not have a significant effect on RMSE and correlation values.

Stone (1992) and Cheng and Yuan (2010), within two-parameter logistic model, found that sample size does not affect error significantly within the estimation of ability parameters.

The result of the study showed that the best estimations for both left-skewed and right-skewed ability parameter distribution was observed in condition which b distribution was normal. Stone (1992) stated that, within right-skewed and normal ability parameter distribution, the best estimations appear in condition that the item difficulty is medium. In addition, he added that the worst estimations appear within easy items. Similarly, in this study, for right-skewed ability parameter distribution, the most defective estimations are made when *b* parameter distribution is right-skewed. Wollack et al. (2002) stated that parameter recovery is best done with the medium-difficulty items and worst done with extreme (easy or difficult) items. Similarly, in this study, Yen (1987) analyzed the conditions in which item difficulty is easy, average and difficult, and worked with 20-items test length, normal ability distribution and with the sample size of 1000. The results of his study revealed that the highest measurement precision was obtained from medium-difficulty items.

Findings about the effect of test length show that, within right-skewed ability distribution and other conditions (normal, uniform, and right-skewed) except for left-skewed item difficulty distribution, measurement precision increases when test length increases. In the literature, there are similar studies in accordance with the relevant results of dichotomous models and polytomous models (Ankenmann & Stone, 1992; Boughton et al., 2001; Hulin et al., 1982; Kieftenbeld & Natesan, 2012; Lautenschlager et al., 2006; Preinerstorfer & Formann, 2012; Roberts & Laughlin, 1996; Seong et al., 1997; Stone, 1992; Swaminathan & Gifford, 1979). For 3PL of dichotomous models Swaminathan and Gifford (1979), Hulin et al. (1982) and for 2PL Stone (1992) identified that measurement precision increase when test length increases. For left-skewed ability distribution, no effect of test length was observed. In the literature, there are studies which the ability estimation of test length do not affect measurement precision (Wollack & Cohen, 1998; Wollack et al., 2002). Wollack et al. (2002) had similar results to this study. They found that the increase of test items from 20 to 30 does not develop $P_{ik}(\theta j)$ estimation.

In this study, in accordance with the results obtained from the individuals who have right-skewed ability parameter, it can be suggested that test developers should ensure that number of items is higher when *b* parameters are distributed normal, uniform or right-skewed, and ensure that number of items is lower when *b* parameters have left-skewed as long as it does not decrease content validity. In addition, as measurement precision will be higher when *b* parameter distribution is normal (independently from ability parameter), it is suggested that *b* parameters in the test should have normal distribution as long as it is relevant with the purpose. In other words, when most of the items have a medium-difficulty level, it would be more appropriate in accordance with the results if difficult and easy items are fewer. Another suggestion for the test developers is that most of the test items should not be very difficult (when *b* parameter distribution is left-skewed) or very easy (when *b* parameter distribution is right-skewed). Because within this kind of *b* parameter distributions, measurement precision may be lower when compared to normal and uniform distribution.

In this study, right-skewed and left-skewed ability parameters were produced from the real data, and conditions were created with reference to different sample size, different *b* parameter distributions and different test lengths. Other researchers can conduct some other studies in other conditions that have estimation method, model, and number of categories for polytomous items, number of replication, estimation program etc. rather than sample size and test length. In addition, they can research the effect of different *b* parameter distributions on measurement precision when ability parameters have normal and uniform distribution. While this study was conducted for dichotomous data, other studies can be conducted for polytomous. Although this study was done using 3-parameter logistic model, other researchers can use other models. In conclusion, while this study analyzed measurement precision within ability parameter estimation, some other studies, within same conditions, can analyze the change of measurement precision within item parameter estimation.

REFERENCES

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*(4), 255-278. Doi: 10.1207/s15324818ame0704_1

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573. doi: 10.1007/BF02293814

Ankenmann, R. D., & Stone, C. A. (1992, April). *A Monte Carlo study of marginal maximum likelihood parameter estimates for the graded model*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bahry, L. M. (2012). *Polytomous item response theory parameter recovery: An investigation of non-normal distributions and small sample size* (Unpublished Master Thesis, University of Alberta Department of Educational Psychology, Edmonton). Retrieved from https://era.library.ualberta.ca/items/55cebca1-82a2-44b5-ab78-aad933bbf147.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153-169. doi: 10.1177/01466216980222005

Bhakta, B., Tennant, A., Horton, M., Lawton, G., & Andrich, D. (2005). Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education*, *5*(1), 9. doi: 10.1186/1472-6920-5-9

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(4), 354-372. doi: 10.21031/epod.346650

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29-51. doi: 10.1007/BF02291411

Boughton, K. A., Klinger, D. A., & Gierl, M. J. (2001, April). *Effects of random rater error on parameter recovery of the generalized partial credit model and graded response model*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Cheng, Y., & Yuan, K. H. (2010). The impact of fallible item parameter estimates on latent trait recovery. *Psychometrika*, *75*(2), 280-291. doi: 10.1007/s11336-009-9144-x

Craig, S. B., & Kaiser, R. B. (2003). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods*, *6*(1), 44-60. doi: 10.1177/1094428102239425

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont CA: Wadsworth group/Thomson learning.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *4*(2), 64-75. Retrieved from http://dergipark.gov.tr/epod/issue/5800/77213.

De Ayala, R. J., & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. *Applied Psychological Measurement*, *23*(1), 3-19. doi: 10.1177/01466219922031130

DeMars, C. E. (2002, April). *Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*(4), 275-288. doi: 10.1177/0146621603027004003

Doğan, N., & Tezbaşaran, A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *25*(25), 58-67. Retrieved from http://dergipark.gov.tr/download/article-file/87861.

Dolma, S. (2009). *Çok ihtimalli Rasch modeli ile derecelendirilmiş yanıt modelinin örtük özellikleri tahminleme performansı açısından simülasyon yöntemiyle karşılaştırılması* (Yayımlanmamış Doktora Tezi, İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul). Erişim adresi: https://tez.yok.gov.tr/UlusalTezMerkezi/.

Fotaris, P., Mastoras, T., Mavridis, I., & Manitsaris, A. (2010, September). Performance evaluation of the small sample dichotomous IRT analysis in assessment calibration. In *Computing in the Global Information Technology (ICCGI), 2010 Fifth International Multi-Conference on* (pp. 214-219). IEEE. doi: 10.1109/ICCGI.2010.19

Guyer, R., & Thompson, N. (2011). *Item response theory parameter recovery using Xcalibre 4.1*. Saint Paul, MN: Assessment Systems Corporation.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory principles and applications* (2. Ed.). USA: Kluwer-Nijhoff.

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

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

Han, K. T., & Hambleton, R. K. (2007). *User's manual: WinGen* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation, 17*(1), 1-24. Retrieved from http://pareonline.net/getvn.asp?v=17&n=1.

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

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi: 10.16986/HUJE.2016015182

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260. doi: 10.1177/014662168200600301

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 36*(5), 399–419. doi: 10.1177/0146621612446170

Koğar, H. (2015). Madde tepki kuramına ait parametrelerin ve model uyumlarının karşılaştırılması: Bir Monte Carlo çalışması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(1), 142-157. doi: 10.21031/epod.02072

Köse, İ. A. (2010). *Madde tepki kuramına dayalı tek boyutlu ve çok boyutlu modellerin test uzunluğu ve örneklem büyüklüğü açısından karşılaştırılması* (Yayınlanmamış Doktora Tezi, Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/.

Lautenschlager, G. J., Meade, A. W., & Kim, S. H. (2006, April). *Cautions regarding sample characteristics when using the graded response model.* Paper presented at the 21st Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* USA: Information Age Publishing.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). Methods and procedures in TIMSS 2015. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi: 10.1007/BF02296272

Montgomery, M., & Skorupski, W. (2012, April). *Investigation of IRT parameter recovery and classification accuracy in mixed format.* Paper presented at the annual meeting of the Nation Council of Measurement in Education, British Columbia.

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–176. doi: 10.1002/j.2333-8504.1992.tb01436.x

OECD. (2017). *PISA 2015 Technical Report*. Paris: PISA, OECD Publishing.

Preinerstorfer, D., & Formann, A. K. (2012). Parameter recovery and model selection in mixed Rasch models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 251-262. doi: 10.1111/j.2044-8317.2011.02020.x

Preston, K. S. J., & Reise, S. P. (2014). Estimating the nominal response model under nonnormal conditions. *Educational and Psychological Measurement*, *74*(3), 377-399. doi: 10.1177/0013164413507063

Reise, S. P., & Yu, J. (1990). Parameter recover in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144. doi: 10.1111/j.1745-3984.1990.tb00738.x

Roberts, J. S., & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*(3), 231-255. doi: 10.1177/014662169602000305

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*(4, Pt. 2), 100.

Seong, T. J., Kim, S. H., & Cohen, A. S. (1997, March). *A comparison of procedures for ability estimation under the graded response model.* Paper presented at the annual meeting of the Nation Council of Measurement in Education, Chicago.

Sen, S. (2014). *Robustness of mixture IRT models to violations of latent normality* (Doctoral dissertation, University of Georgia, Athens). Retrieved from http://tez.yok.gov.tr/ UlusalTezMerkezi/.

Sen S., Cohen A.S., Kim SH. (2015) Robustness of Mixture IRT Models to Violations of Latent Normality. In: Millsap R., Bolt D., van der Ark L., Wang WC. (eds) Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics, vol 89. Springer, Cham. doi: 10.1007/978-3-319-07503-7_3

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two- parameter logistic response model: An evaluation of MULTILOG. A*pplied Pscyhological Measurement, 16*(1), 1-16. doi: 10.1177/014662169201600101

Swaminathan, H. & Gifford, J. A. (1979, April). *Estimation of parameters in the three- parameter latent trait model*. Paper presented at the annual meeting of AERA-NCME, San Francisco.

Tate, R. (2000). Robustness of the school-level polytomous IRT model. *Educational and Psychological Measurement, 60*(1), 20-37. doi: 10.1177/00131640021970349

Thissen, D., Chen, W. H. & Bock, D. (2003). *MULTILOG 7.03*. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397-412. doi: 10.1007/BF02293705

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9. Retrieved from http://pareonline.net/getvn.asp?v=16&n=1.

Wang, W. C., & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement, 65*(3), 376-404. doi: 10.1177/0013164404268673

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y. S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological measurement, 26*(3), 339-352. doi: 10.1177/0146621602026003007

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*(2), 144-152. doi: 10.1177/01466216980222004

Yavuz, G., & Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement, 77*(2), 263-274. doi: 10.1177/0013164416661220

Farklı Yetenek Dağılımlarında Madde Güçlük Dağılımı, Test Uzunluğu ve Örneklem Büyüklüğünün İncelenmesi

Giriş

Madde tepki kuramının (MTK) karakteristik özellikleri sayesinde bireye uygun test geliştirme, madde yanlılığını belirleme, testleri eşitleme gibi durumlarda ilerleme sağlanmış, sınırlılıklar giderilmiştir (Hambleton ve Swaminathan, 1985). MTK'nın birçok avantajından dolayı PISA, TIMSS gibi uluslararası sınavlarda tercih edildiği görülmektedir. Ayrıca ulusal ve uluslararası birçok araştırmada sınavlardan elde edilen sonuçların MTK bağlamında değerlendirildiği de görülmektedir. Bireyler için oldukça önemli bir konu olan ve eğitimde kullanılan sınavlar farklı amaçlarla hazırlanmaktadır. Bu amaçlar arasında öğrencileri seçme ve yerleştirme, düzey belirleme, girdi özelliklerini belirleme, öğrencileri sıralama vb. yer alabilir. Sınavlar hazırlanış ve uygulanış amacına veya testi alan bireylerin özelliklerine ve /veya sayısına göre farklı psikometrik özelliklere de sahip olacaktır. Örneğin bir testi alan birey sayısının fazla fakat test sonucu ile karar verilecek birey sayısı az ise hazırlanan testin zor olması beklenen bir durumdur. Ancak seçme ve yerleştirme amacından çok bireylerin var olan bilgilerinin tespiti için hazırlanan bir sınavın ise seçme ve yerleştirme sınavına göre daha kolay olması hatta mümkünse çoğunluğunun orta güçlükte maddelerden oluşması daha istendik bir durumdur. Burada asıl olan testlerde ölçme ve değerlendirme açısından sağlanması gereken geçerlik ve güvenirliğin bu durumdan nasıl etkileneceğinin belirlenmesidir. Ayrıca testi alan bireylerin yetenek dağılımlarının farklılaşmasının da geçerlik ve güvenirliğe olan etkisinin belirlenmesi de önemlidir.

Bu çalışmada ulusal bir sınavdan elde edilen parametrelere dayanarak birey dağılımının sağa ve sola çarpık olması durumunda, farklı b parametresi dağılımlarının, test uzunluğunun ve örneklem büyüklüğünün birey parametresi kestiriminde ölçme kesinliğine etkisi incelenmiştir. Literatürde

birey dağılımı türü, örneklem büyüklüğü ve test uzunluğu koşullarının ölçme kesinliğine etkisinin incelendiği sıklıkla görülmektedir. Ancak farklı birey dağılımları, test uzunlukları ve örneklem büyüklüklerinde farklı *b* parametresi dağılımlarının ölçme kesinliğine etkisinin incelendiği çalışmalara literatürde rastlanmamıştır. Burada farklı madde güçlüğü dağılımlarına dayalı olarak türetilen dört farklı testin işe koşulması çalışmanın ayrıca önemini oluşturmaktadır.

1. Sağa çarpık yetenek dağılımında, farklı test uzunlukları, örneklem büyüklükleri ve madde güçlük dağılımlarının yetenek parametresi kestiriminin ölçme kesinliğine etkisi nedir?
2. Sola çarpık yetenek dağılımında, farklı test uzunlukları, örneklem büyüklükleri ve madde güçlük dağılımlarının yetenek parametresi kestiriminin ölçme kesinliğine etkisi nedir?

Yöntem

Araştırma kapsamında kullanılan koşulların oluşturulması amacıyla veriler üretildiğinden bu çalışma simülasyon çalışmasıdır. Araştırmada öncelikle birey parametreleri elde edilmiştir. Bu amaçla, liselere geçişte uygulanan ulusal öğrenci seçme sınavının 20 maddelik matematik alt testinden elde edilen veriler kullanılmıştır. Araştırmada 500, 1000, 2500, 5000 ve 10000 olmak üzere toplam beş örneklem büyüklüğü belirlenmiştir. Simülasyon çalışması için ilk aşamada gerçek birey parametreleri elde edilmiştir. Sağa çarpık birey parametrelerinin elde edilmesinde her bir örneklem büyüklüğü için gerçek veriden random gruplar seçilmiştir. Sola çarpık birey parametrelerinin elde edilmesinde ise verinin tamamından kasıtlı örnekleme yoluyla çarpıklık ≈-1,00 olacak şekilde her örneklem büyüklüğünde veri setleri seçilmiştir. Simülasyonun 2. aşamasında ise madde parametreleri türetilmiştir. Bu aşamada farklı b parametresi dağılımına sahip (normal dağılım, tekdüze dağılım, sola çarpık ve sağa çarpık dağılım) hem 20 maddelik hem de 30 maddelik testler oluşturulmuştur. Madde parametrelerinin üretilmesinde a parametre değeri min= 0,5 maks=2 olarak, c parametre değeri min= 0 maks=0,05 olarak belirlenmiştir. Sola çarpık b parametresi dağılımı için α=8; β=2; sağa çarpık b parametresi dağılımı için α=2; β=8; tekdüze b parametre dağılımı için min=-3; maks=+3; normal b parametresi dağılımı için ort=0; Ss= 1 değerleri kullanılarak araştırma kapsamında kullanılacak dört ayrı madde güçlüğü dağılımı oluşturulmuştur.

Araştırma kapsamına alınan 80 koşul (2 birey dağılımı x 5 örneklem büyüklüğü x 4 b parametresi dağılımı x 2 test uzunluğu) Wingen 3 programı (Han, 2007) yardımıyla oluşturulmuştur. MTK'de parametre iyileştirme çalışmalarında genel olarak ölçme kesinliği hesaplaması yapılmaktadır. Ölçme kesinliğini incelemek amacıyla "hata kareleri ortalamasını karekökü" (Root Mean Squared Error (RMSE)) ve "ortalama mutlak farklılık" (Absolute Average Deviation (AAD)) değerleri hesaplanmıştır.

1. Alt probleme ilişkin bulgular: Sağa çarpık birey dağılımında ele alınan tüm örneklem büyüklüklerinde ölçme kesinliği en yüksek; b parametresi dağılımı normal ve test uzunluğu 30 madde olduğunda, en düşük ise b parametresi sağa çarpık ve test uzunluğu 20 madde olduğunda elde edilmiştir. Ayrıca ölçme kesinliğinin normal b dağılımdan sonra en yüksek tekdüze b dağılımında olduğu gözlemlenmiştir. Araştırmanın sonuçları test uzunluğu açısından incelendiğinde ise, normal, tekdüze ve sağa çarpık b dağılımlarında genel olarak 20 maddelik teste ilişkin ölçme kesinliğinin 30 maddelik teste göre daha düşük olduğu belirlenmiştir. Bu b dağılımlarının aksine sola çarpık b dağılımda ise 20 maddelik testin ölçme kesinliğinin 30 maddelik teste göre daha yüksek olduğu görülmüştür. Sonuç olarak test uzunluğu arttıkça ölçme kesinliğinin de arttığı belirlenmiştir. Son olarak örneklem büyüklüğünün birey parametresinin kestiriminde ölçme kesinliğine önemli bir etkisinin olmadığı gözlemlenmiştir.
2. Alt probleme ilişkin bulgular: Sola çarpık birey dağılımında ele alınan farklı test uzunluklarında ve örneklem büyüklüklerinde b parametresi dağılımı normal olduğunda ölçme kesinliğinin en yüksek düzeyde olduğu ve bunu tekdüze dağılımın takip ettiği

söylenebilir. Ayrıca en düşük ölçme kesinliğinin de tüm test uzunluğu ve örneklem büyüklüklerinde en düşük sola çarpık b dağılımında olduğu görülmüştür. Son olarak sola çarpık birey dağılımı için örneklem büyüklüğünün ve test uzunluğunun birey parametrelerinin kestirim üzerinde önemli bir etkisi olmadığı gözlemlenmiştir.

Sonuç ve Tartışma

Araştırmadan elde edilen sonuçlarda, hem sağa hem de sola çarpık birey dağılımda farklı b dağılımına sahip her bir test için örneklem büyüklüğü arttıkça ölçme kesinliği için hesaplanan RMSE ve, AAD değerlerinde çok fazla değişim olmadığı görülmüştür. Sağa çarpık birey dağılımı için tüm örneklem büyüklüklerinde test uzunluğunun etkisi incelendiğinde ise test uzunluğu arttığında RMSE ve AAD değerlerinin genel olarak azaldığı gözlemlenmiştir. Ancak sola çarpık birey dağılımı için test uzunluğundaki değişimin ölçme kesinliğini önemli derecede etkilemediği görülmüştür. Ayrıca sağa ve sola çarpık birey dağılımlarında, tüm örneklem büyüklüğü ve test uzunlukları için; en yüksek ölçme kesinliği b parametresi dağılımı normal olduğunda elde edilmiştir. Normal b dağılımını ise b parametresinin tekdüze dağıldığı koşul izlemiştir. Son olarak sağa çarpık birey dağılımı için RMSE ve AAD değerlerinin en yüksek sağa çarpık *b* dağılımında olduğu, sola çarpık birey dağılımda ise en yüksek sola çarpık b dağılımında olduğu gözlemlenmiştir.

Araştırmanın sonuçları doğrultusunda test geliştiricilere sola çarpık *b* parametre dağılımı yani maddelerin çoğunluğunun zor olması ya da sağa çarpık *b* parametre dağılımı yani maddelerin çoğunluğunun kolay olması önerilmez. Çünkü bu tip *b* parametresi dağılımlarında ölçme kesinliği normal ve tekdüze *b* parametresi dağılımına kıyasla daha düşük elde edilebilmektedir. Başka araştırmalarda örneklem büyüklüğü ve test uzunluğu yerine kestirim yöntemi, model, çoklu puanlanan maddeler için kategori sayısı, tekrar sayısı, kestirim programı vb. gibi koşulların ölçme kesinliğine etkisi incelenebilir. Ayrıca yetenek parametreleri normal ve tekdüze dağılıma sahip olduğunda, farklı *b* parametresi dağılımlarının ölçme kesinliğine etkisi de araştırılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2018; 9(3);295-311

Investigating The Effect of Exposure-Control Strategies on Item Selection Methods in MCAT

Xiuzhen MAO * Burhanettin ÖZDEMİR ** Yating WANG*** Tao XIN****

Abstract
This study aims to investigate the effect of different item exposure controlling strategies on item selection methods in the context of multidimensional computerized adaptive testing (MCAT). Additionally, this study aims to examine to what extend the restrictive threshold (RT) and the restrictive progressive (RPG) exposure methods suppress the item exposure rates and increase the exposure rates of underexposed items without losing psychometric precision in MCAT. For this purpose, the performance of four item selection methods with and without exposure controls are evaluated and compared so as to determine how results differ when item exposure controlling strategies are applied with Monte-Carlo simulation method. The four item selection methods employed in this study are D-optimality, Kullback–Leibler information (KLP), the minimized error variance of linear combination score with equal weight (V1), the composite score with optimized weight (V2). On the other hand, the maximum priority index (MPI) method proposed for unidimensional CAT and two other item exposure control methods, that are RT and RPG methods proposed for cognitive diagnostic CAT, are adopted. The results show that: (1) KLP, D-optimality, and V1 performed better in recovering domain scores, and all outperformed V2 with respect to precision; (2) although V1 and V2 offer improved item bank usage rates, KLP, D-optimality, V1, and V2 produced an unbalanced distribution of item exposure rates; (3) all exposure control strategies improved the exposure uniformity greatly and with very little loss in psychometric precision; (4) RPG and MPI performed similarly in exposure control, and outperformed RT exposure control method.

Keywords: Multidimensional computerized adaptive testing, item selection methods, exposure control strategies.

INTRODUCTION

The fact that test items are chosen sequentially and adaptively in computerized adaptive testing (CAT) has broken the traditional testing mode in which thousands of people respond to the same items at the same time. Nowadays, CAT is increasingly favored by test practitioners and researchers for its higher efficiency, shorter test time, and lower pressure compared to paper and pencil (P&P) testing. Another more fascinating characteristic of CAT is that different item response models can be applied, including unidimensional, multidimensional, and cognitive diagnostic models.

Multidimensional computer adaptive testing (MCAT) possesses the advantages of both multidimensional item response theory (MIRT) and CAT. On the one hand, a large number of studies based on different test conditions have declared that MCAT provides higher efficiency than unidimensional CAT. For example, Segall (1996) employed simulated data based on nine adaptive power tests of the Armed Services Vocational Aptitude Battery (ASVAB) to show that MCAT reduced by about one-third the number of items required to generate equal or higher reliability with similar precision to unidimensional CAT. Luecht (1996) demonstrated that MCAT can reduce the number of items for tests with content constraints by 25–40%. Further, Wang and Chen (2004)

* Assistant Prof. Dr., Sichuan Normal University, Sichuan-China, maomao_wanli@163.com, ORCID ID: 0000-0001-8245-3633
** Assistant Prof. Dr., Siirt University, Faculty of Education, Siirt-Turkey, b.ozdemir025@gmail.com, ORCID ID: 0000-0001-7716-2700
*** Prof. Dr., Sichuan Normal University School of Education, Sichuan-China, 1358178364@qq.com, ORCID ID: 0000-0001-9328-5380
**** Prof. Dr., Beijing Normal University, Institute of Educational Statistics and Measurement, Beijing-China, xintao@bnu.edu.cn, ORCID ID: 0000-0003-2297-2604

DOI: *10.21031/epod.407521*
Received: 21.03.2018
Accepted 30.08.2018

illustrated the higher efficiency of MCAT compared with unidimensional CAT under different latent trait correlations, latent numbers, and scoring levels. On the other hand, the fact that several ability profiles are estimated simultaneously indicates the ability of MCAT to offer detailed diagnostic information regarding domain scores and overall scores. The advantages of multi-dimensionality and high efficiency make MCAT better suited to real tests than unidimensional CAT. Hence, many studies on MCAT have considered real item banks, such as Terra Nova (Yao, 2010), American College Testing (ACT) (Veldkamp & van der Linden, 2002), and ASVAB (Segall, 1996; Yao, 2012, 2014a).

Since Bloxom and Vale (1987) extended unidimensional CAT to MCAT, it has received increasing attention, and several breakthroughs have been reported in the last decade. Among the studies on ability estimation methods, the testing stopping rule, and item replenishing, item selection rules have become popular because of their important role in affecting the test quality and psychometric precision. Thus, most researchers focus on proposing new item selection indices to decrease errors in ability estimation. However, Yao (2014a) pointed out that most item selection methods tend to select a particular type of item, leading to the problem of unbalanced item utility. She also gave an example of the Kullback–Leibler index, which prefers items that have either a high discriminator at each dimension or significantly different discriminators among different dimensions. As another example, the D-optimality index tends to select items with a high discrimination in only one dimension (Wang, Chang, & Boughton, 2011). Nowadays, CAT is increasingly used in many kinds of tests. Hence, item exposure control is important in the application of MCAT, especially for its application to high-stakes tests. Furthermore, few studies have investigated this problem in MCAT. Hence, the goal of the present study is to examine the performance of some exposure control techniques along with item selection methods in MCAT.

To date, many of the exposure control methods used in unidimensional CAT have been generalized to MCAT. For example, Finkelman, Nering and Roussos (2009) extended the Sympson–Hetter (S-H) (Sympson & Hetter, 1985) and Stocking–Lewis (S-L) (Stocking & Lewis, 1998) methods to MCAT. They found that all the S-H, generalized S-H, and generalized S-L methods do well in controlling the maximum item exposure rates. However, simulation experiments to create the exposure control parameters are time-consuming. Furthermore, there still exist some underexposed items. In addition, Yao (2014a) compared S-H with the fix-rate procedure. The fix-rate procedure is similar to the maximum priority index (MPI) method proposed by Cheng and Chang (2009) for unidimensional CAT. She showed that the S-H method performs better in terms of test precision, whereas the latter gives a higher item bank usage and controls the maximum item exposure rate well.

The $|a_{j1}-a_{j2}|$-stratification method (Lee, Ip, & Fuh, 2008) is based on the principle of the a-stratification method (Chang & Ying, 1999). The item bank is stratified according to the absolute value of $a_{j1}-a_{j2}$, where $a=(a_{j1},a_{j2})$ denotes the item discrimination vector of item j. It was reported that the $|a_{j1}-a_{j2}|$-stratification method is effective in combating overused items and increasing the item bank usage. However, this method cannot guarantee that no items are overexposed. Thus, Huebner, Wang, Quinlan, and Seubert (2015) combined $|a_{j1}-a_{j2}|$-stratification with the item eligibility method (van der Linden & Veldkamp, 2007) with the aim of enhancing the balance of item exposure. This combination method improves the exposure rates of underused items and suppresses the observed maximum item exposure rate. However, these two methods are restricted to tests with two dimensions. Constructing a suitable functional of the discrimination parameter for tests with more than two dimensions remains an important research problem.

It is well known that the uniformity of item exposure rates is affected by the numbers of overexposed and underexposed items. Of the above mentioned exposure control methods used in MCAT, the S-H, generalized S-H, generalized S-L, fix-rate, and item eligibility methods perform well in suppressing the maximum item exposure rates, and the $|a_{j1}-a_{j2}|$-stratification method effectively improves the

utility of underexposed items. Although the combination method used by Huebner, et al. (2015) performs well in both aspects, it is only suitable for tests with two dimensions.

The uniformity of item exposure rates and measurement precision are the two most important considerations during the application of MCAT to practical tests, especially for high-stakes tests. Because they always trade-off with one another, practitioners hope to find some item selection method that not only guarantees test precision, but also decreases the maximum item exposure rate while increasing the exposure rate of underexposed items. However, there are no methods that can effectively balance item exposure rates for tests with more than two dimensions. In addition, there are two other exposure control methods that have not been studied for MCAT: the restrictive threshold (RT) method and the restrictive progressive (RPG) method. It has been reported that they perform well in balancing the item exposure rate of cognitive diagnostic CAT (Wang, Chang, & Huebner, 2011). Therefore, the focus of the present study is whether RT and RPG can simultaneously suppress the maximum item exposure rates and increase the exposure rates of underexposed items without losing psychometric precision in MCAT. Further, their performance is compared with that of the MPI method.

METHOD

A Monte Carlo simulation study was conducted to evaluate and compare the effectiveness of the above exposure control methods. Matlab (version7.10.0.499) was used to write MCAT codes and run the simulation conditions.

Design of Simulation Study

Item bank construction: Although Stocking (1994) suggests that the pool should contain at least 12 times as many items as the test length, many simulation studies on MCAT have used a more restrictive item bank. For example, the item bank used by van der Linden (1999) contained 500 items while the test length was 50; Lee, et al. (2008) used an item bank of 480 items with test lengths of 30 and 60; and the item banks described in Veldkamp and van der Linden (2002) and Mulder and van der Linden (2009) contained fewer than 200 items while the test length was greater than 30. Thus, it is reasonable to construct an item bank of 450 items for a test length of 30.

To simplify the experimental conditions, most simulation studies generate item parameters and item responses according to M-2PL or M-3PL with the assumption that there are two or three dimensions (van der Linden, 1999; Veldkamp & van der Linden, 2002; Lee et al., 2008; Mulder & van der Linden, 2009; Finkelman et al., 2009; Wang, Chang, & Boughton, 2013; Wang & Chang, 2011). Hence, without loss of generality, the items in our simulation contained three dimensions, and the item parameters of the M-2PL model were generated in a similar way to those of Yao and Richard (2006) and Wang and Chang (2011). Specifically, (a_{j1}, a_{j2}, a_{j3}) for item $j(j = 1,2,...450)$ were drawn from $\log N(0, 0.5)$ independently and $b_j (j = 1,2,...450)$ were drawn from $N(0,1)$ and each condition is replicated for 100 times.

Examinees and item responses: All 5000 examinees were simulated uniformly from a multivariate normal distribution, as in previous researches (Wang & Chang, 2011; Yao, Pommerich, & Segall, 2014; Wang et al., 2013). Three levels of correlation were considered in the experiments. The mean ability was [0, 0, 0] and the variance-covariance matrix was:

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix} \quad (\rho = 0.3, 0.6, 0.8)$$

Let P_{ij} and x_{ij} denote the correct response probability and actual response (0 or 1) corresponding to the *j*th $(j = 1, 2, ..., 450)$ item and the *i*th $(i = 1, 2, ..., 5000)$ examinee. P_{ij} was computed from the M-2PL model, and u_{ij} was selected uniformly from (0, 1). We set x_{ij} = 1 if $P_{ij} \geq u_{ij}$. Otherwise, if $P_{ij} < u_{ij}$, x_{ij} = 0.

Item selection methods: Four item selection methods with and without the three exposure control methods yields a total of 16 item selection methods.

Estimation of ability: The initial abilities were selected from the standard multivariate normal distribution. MAP was used to update the domain abilities during the test, and multivariate standardized normality was applied as the prior distribution.

Evaluation criteria: The bias and mean square error (MSE) of each dimension were used to evaluate the precision of the ability estimations. The formula for bias and MSE are as follows:

$$Bias_l = \frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{\theta}_l - \theta_l) \qquad (l = 1,2,3), \tag{1}$$

$$MSE_l = \frac{1}{N} \cdot \sum_{i=1}^{N} (\hat{\theta}_l - \theta_l)^2 \qquad (l = 1,2,3). \tag{2}$$

To assess the effect of exposure rates, we used (a) the number of items never administered and the number of items with exposure rates greater than 0.2, (b) the χ^2 statistic, and (c) the test overlap rate. The formula χ^2 statistic is as follows:

$$\chi^2 = \sum_{i=1}^{N} \frac{(er_i - \overline{er})^2}{\overline{er}}. \tag{3}$$

Smaller values of χ^2 indicate smaller differences between the observed and expected item exposure rates. Finally, the test overlap rate was computed according to the expression proposed by Chen, Ankenmann, and Spray (2003):

$$\hat{T} = \frac{M}{L} S_{er}^2 + \frac{L}{M}. \tag{4}$$

where S_{er}^2 denotes the variance of item exposure rates . Generally, smaller values of $\hat{T}$ demonstrate more balanced item utility.

In the following sections, we first introduce the MIRT model employed in this study and the ability estimation method. Then, some item selection indices and exposure control strategies are described. The performance of four item selection indices with and without each of the three exposure control strategies under different latent trait correlation levels are examined through a series of simulation experiments. The results, conclusions, and discussion are given in the final two sections.

MIRT Model and Ability Estimation Method

Multidimensional Two-Parameter Logistic (M-2PL) Model

MIRT models are usually classified as compensatory or non-compensatory based on whether a strong ability can compensate for other weak profiles. Bolt and Lall (2003) reported that both types are able to fit the data generated by non-compensatory models, but non-compensatory models cannot

match the data generated from compensatory models. Thus, because of the advantages of compensatory models and the wide usage of MCAT in dealing with dichotomous items (van der Linden, 1999; Veldkamp & van der Linden, 2002; Mulder & van der Linden, 2010), the M-2PL model was adopted to simulate item parameters and generate item responses.

For some item j, M-2PL includes a scalar difficulty parameter b_j and discrimination vector $a_j = (a_{j1}, a_{j2}, ..., a_{jD})^T$ (McKinley & Reckase, 1982), where *T* denotes the transpose and *D* is the number of dimensions. For an examinee with ability $\theta = (\theta_1, \theta_2, ..., \theta_D)^T$, the item response function can then be described as:

$$P_j(\vec{\theta}) = P(x_j = 1 | \vec{\theta}, \vec{a}_j, b_j) = \frac{1}{1 + \exp[-(\vec{a}_j^T \cdot \vec{\theta} - b_j)]}. \quad (5)$$

where $\vec{a}_j^T \cdot \vec{\theta} - b_j = \sum_{l=1}^{D} a_{jl} \cdot \theta_l - b_j$ denotes a straight line in *D*-dimensional space. The compensatory features of M-2PL originate from the fact that all examinees giving equal $\vec{a}_j^T \cdot \vec{\theta}$ possess the same response probability.

Ability Estimation Method: Maximum a Posteriori (MAP) Estimation

In this study, MAP is adopted for its competitive precision and easier computation compared to expected a posteriori (EAP) ability estimation method in MIRT. Yao (2014b) compared MAP, expected a posteriori (EAP), and maximum likelihood estimation (MLE) in a simulation experiment using item parameters estimated from the ASVAB Armed Forces Qualification Test. She pointed out that: (a) MLE generates smaller bias and larger root mean square error (RMSE), whereas MAP and EAP using strong prior information or standard normal priors produced higher precision in the recovery of ability, while EAP estimation takes a longer time than MAP. Recently, Huebner, et al. (2015) compared EAP with MLE in MCAT, and proved that EAP always produces more stable results and lower mean square error in the ability estimators than MLE.

Let $f(\vec{\theta})$ denote the prior density function of $\vec{\theta}$. This is assumed to be a multivariate normal distribution with mean value $\vec{\mu}_0$ and variance-covariance matrix Σ_0. For convenience, the response to item j is indicated as x_j, and $\vec{X}_{k-1}$ represents the response vector of the first $k-1$ items administered. The posterior density function of $\vec{\theta}$ is denoted by $f(\vec{\theta} | \vec{X}_{k-1})$. Based on Bayes' theorem, $f(\vec{\theta} | \vec{X}_{k-1}) \propto L(\vec{X}_{k-1} | \vec{\theta}) \cdot f(\vec{\theta})$, where $L(\vec{X}_{k-1} | \vec{\theta})$ denotes the likelihood function. Hence, the goal of MAP is to find the mode that maximizes the posterior density function $f(\vec{\theta} | \vec{X}_{k-1})$. That is, the ability estimator $\hat{\vec{\theta}}^{MAP}$ is equivalent to the solution of $\frac{\partial \log f(\vec{\theta} | \vec{X}_{k-1})}{\partial \vec{\theta}_l} = 0 \quad (l = 1,2,...,D)$. Furthermore, Newton-Raphson iteration can be used to solve this equation (for more details see, Yao, 2014b).

Item Selection Methods and Exposure Control Strategies

To simplify the description, we first introduce some notation. N represents the number of examinees, and L is the test length. Set R refers to the item bank, which has a capacity of M. Set

$R_{k-1} = R \setminus \{i_1, i_2, ...i_{k-1}\}$ and $\hat{\vec{\theta}}^{k-1}$ express the remainder of the item bank and the temporary estimator after administering the first $k-1$ items, respectively.

Item Selection Methods

The following four indices are chosen as item selection criteria based on the consideration of computation complexity and running time.

D-optimality: The Fisher information of each item in MIRT is no longer a number, but a matrix. Specifically, the Fisher information for the *j*th item in M-2PL is

$$I_j(\vec{\theta}) = P_j(\vec{\theta}) \cdot (1 - P_j(\vec{\theta})) \cdot (\vec{a}_j^T \vec{a}_j). \tag{6}$$

After $k-1$ items have been administered, the estimators form an ellipse or sphere V_{k-1}. To decrease the size or volume of V_{k-1} as quickly as possible, Segall (1996) proposed that the *k*th item should maximize the determinant of the posterior test Fisher information matrix. Thus, the Bayesian item selection rule is expressed as

$$D_k = \max\{|\ I_{k-1}(\hat{\vec{\theta}}^{k-1}) + I_j(\hat{\vec{\theta}}^{k-1}) + \Sigma_0^{-1}|, \qquad j \in R_{k-1}\}. \tag{7}$$

where $I_{k-1}(\hat{\vec{\theta}}^{k-1})$ represents the test information of the first $k-1$ items already be administered calculated at the current estimated ability, and $I_j(\hat{\vec{\theta}}^{k-1})$ indicates the Fisher information of the *j*th $(j \in R_{k-1})$ candidate item. This method was called D-optimality by Mulder and van der Linden (2009), and the item with the largest D_k is chosen from the remainder pool.

Posterior expected Kullback–Leibler information (KLP): This method is obtained by weighting the KL information according to the posterior distribution of ability. That is, the *k*th item is selected according to

$$KLP_k = \max\{\int_{\vec{\theta}} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) \cdot f(\vec{\theta} \mid \vec{X}_{k-1}) d\vec{\theta}, \qquad j \in R_{k-1}\}. \tag{8}$$

where

$$\begin{aligned} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) &= E_{\vec{\theta}} \log[\frac{P_j(x_j \mid \vec{\theta}, \vec{a}_j, b_j)}{P_j(x_j \mid \hat{\vec{\theta}}^{k-1}, \vec{a}_j, b_j)}] \\ &= P_j(\vec{\theta}) \log \frac{P_j(\vec{\theta})}{P_j(\hat{\vec{\theta}}^{k-1})} + (1 - P_j(\vec{\theta})) \log \frac{(1 - P_j(\vec{\theta}))}{(1 - P_j(\hat{\vec{\theta}}^{k-1}))}. \end{aligned} \tag{9}$$

The integral interval is generally narrowed to simplify the computation, and (9) is replaced with

$$KLP_k = \max\{\int_{\theta_1^{k-1} - \gamma_j}^{\theta_1^{k-1} + \gamma_j} ... \int_{\theta_D^{k-1} - \gamma_j}^{\theta_D^{k-1} + \gamma_j} KL_j(\hat{\vec{\theta}}^{k-1}, \vec{\theta}) \cdot f(\vec{\theta} \mid \vec{X}_{k-1}) d\theta_1 \cdots d\theta_D, \qquad j \in R_{k-1}\}, \tag{10}$$

where γ_j usually takes a value of $3 / \sqrt{j}$.

Minimum error variance of the linear combination score with equal weight (V1): From the perspective of error variance, van der Linden (1999) suggested that the *k*th item should minimize the error variance of the composite score $\vec{\theta}_\alpha = \sum_{l=1}^{D} \theta_l \cdot w_l$. Let $SEM(\vec{\theta}_\alpha)$ denote the standard error of

measurement (SEM) for composite score $\vec{\theta}_\alpha$. Yao (2012) derived the formula $SEM(\vec{\theta}_\alpha) = (V(\vec{\theta}_\alpha))^{1/2} = (\vec{w}V(\vec{\theta})\vec{w}^T)^{1/2}$, where $V(\vec{\theta})$ is usually approximated by $I_{k-1}(\hat{\vec{\theta}}^{k-1})^{-1}$.

Given equal weights $w = (1/D,\ 1/D,...,\ 1/D)$ among the different dimensions, the item that minimizes $SEM(\vec{\theta}_\alpha)$ will be selected by V1.

Minimum error variance of the linear combination score with optimized weight (V2): The weight that minimizes the SEM of the composite ability is named the optimal weight. Yao (2012) proved the existence of the optimized weight, and derived its formula as:

$$w = \frac{1}{\sum_{o=1}^{D}\sum_{l=1}^{D} b_{ol}} \cdot [1,1,...,1]_{1\times D} \cdot I_{k-1}(\vec{\theta}) \quad . \tag{11}$$

In this expression, b_{ol} denotes the element of $I_{k-1}(\vec{\theta})$ located on the oth row and lth column. The procedure of V2 involves finding the optimal weight vector, then calculating SEM for each candidate item according to the optimal weight. Finally, the item with the lowest SEM is selected from the remainder pool. Note that the optimal weight is updated after administering each item. Thus, the only difference between V2 and V1 is in the determination of the weight used to compute $SEM(\vec{\theta}_\alpha)$.

Item Exposure Controlling Methods

The RT and RPG methods proposed by Wang, et al. (2011) are two exposure control methods used in cognitive diagnostic CAT. Both can be easily generalized to MCAT.

The RT method: In the RT method, a shadow item bank is constructed at the beginning of each test by removing all overexposed items from the original item bank. Each item is then selected at random from the candidate item set constructed beforehand. Let "Index" denote the value of the item selection indices. The candidate item set includes all items whose information values lie in $[\max(Index) - \delta, \max(Index)]$ for both D-optimality and KLP or $[\min(Index), \min(Index) + \delta]$ for V1 and V2. The constant δ is defined as $\delta = [\max(Index) - \min(Index)] \cdot (1 - k/L)^{\beta}$. Larger values of β give a shorter information interval length. As a result, the measurement precision is improved by decreasing the uniformity of the item exposure distribution. In summary, β is used to balance the requirements of item exposure rate control and measurement precision. In this study, β = 0.5 is favored.

The RPG method: The kth (k = 1, 2, …, L) item is selected according to formula (12) for D-optimality and KLP, and according to formula (13) for V1 and V2. These two formulas are as follows:

$$i_k = \max\{(1 - er_j / r^{\max}) \cdot [(1 - k/L)u_j + Index_j \times \beta k/L], \qquad j \in S_{k-1}\} \tag{12}$$

$$i_k = \max\{(1 - er_j / r^{\max}) \cdot [(1 - k/L)R_j + (C - Index_j) \times \beta k/L], \qquad j \in S_{k-1}\}, \tag{13}$$

where er_j denotes the observed exposure rate of item j and $r^{\max}$ denotes the allowed maximum exposure rate. Let H^* be the maximum item information in S_{k-1}. Then, u_j is uniformly extracted from interval $(0, H^*)$. The parameter β plays the same role and takes the same value as in the RT method. The constant C should be greater than all the SEMs; in this study, we set C = 10000. Note

that SEM is always very large for the first several items, and decreases rapidly to less than 1000. Thus, it is better to set *C* to be greater than 1000.

The maximum priority index method (MPI): According to Cheng and Chang (2009), the priority index (*PI*) of item j with the requirement of the maximum exposure rate is expressed as

$$PI_j = \frac{r^{\max} - n_j / N}{r^{\max}} \cdot Index_j, \quad (14)$$

where n_i represents the administration frequency of item j, and "*index* " refers to the D-optimality or KLP index. Finally, the task of the MPI method is to identify the item with the largest *PI*. The role of *C* is similar to that in RPG. For V1 and V2, PI_j should be changed accordingly, that is

$$PI_j = \frac{r^{\max} - n_j / M}{r^{\max}} \cdot (C - Index_j). \quad (15)$$

RESULTS

Results of Ability Estimation

The ability estimations obtained from different MCAT algorithms were compared with respect bias and MSE statistics. Figure 1 depicts mean bias of the three ability dimensions under each item selection method and item exposure control methods with differing correlation between dimensions.

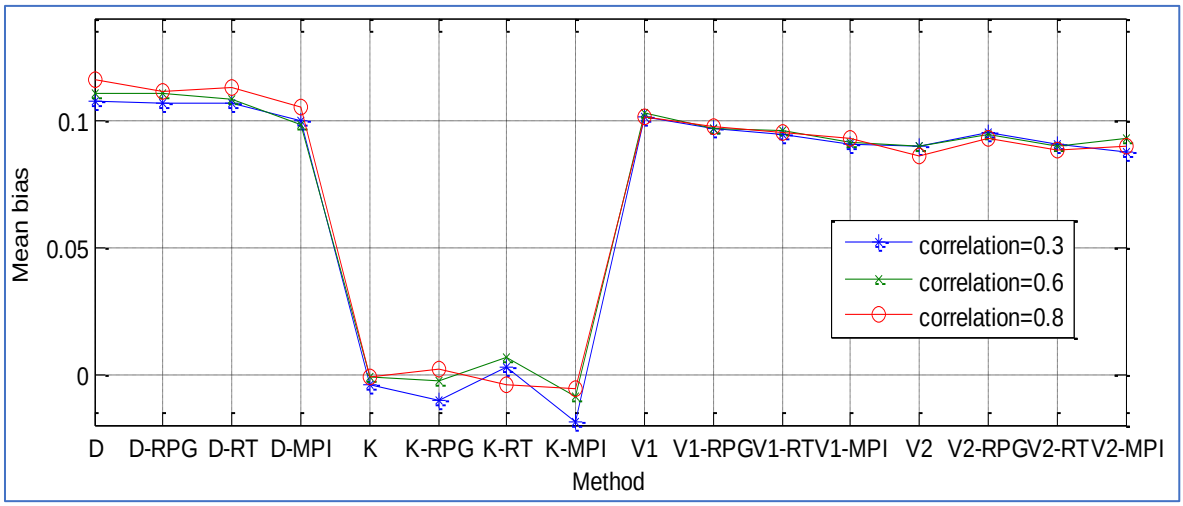


Figure 1. Mean Bias of the Three Ability Dimensions Under Each Item Selection Method

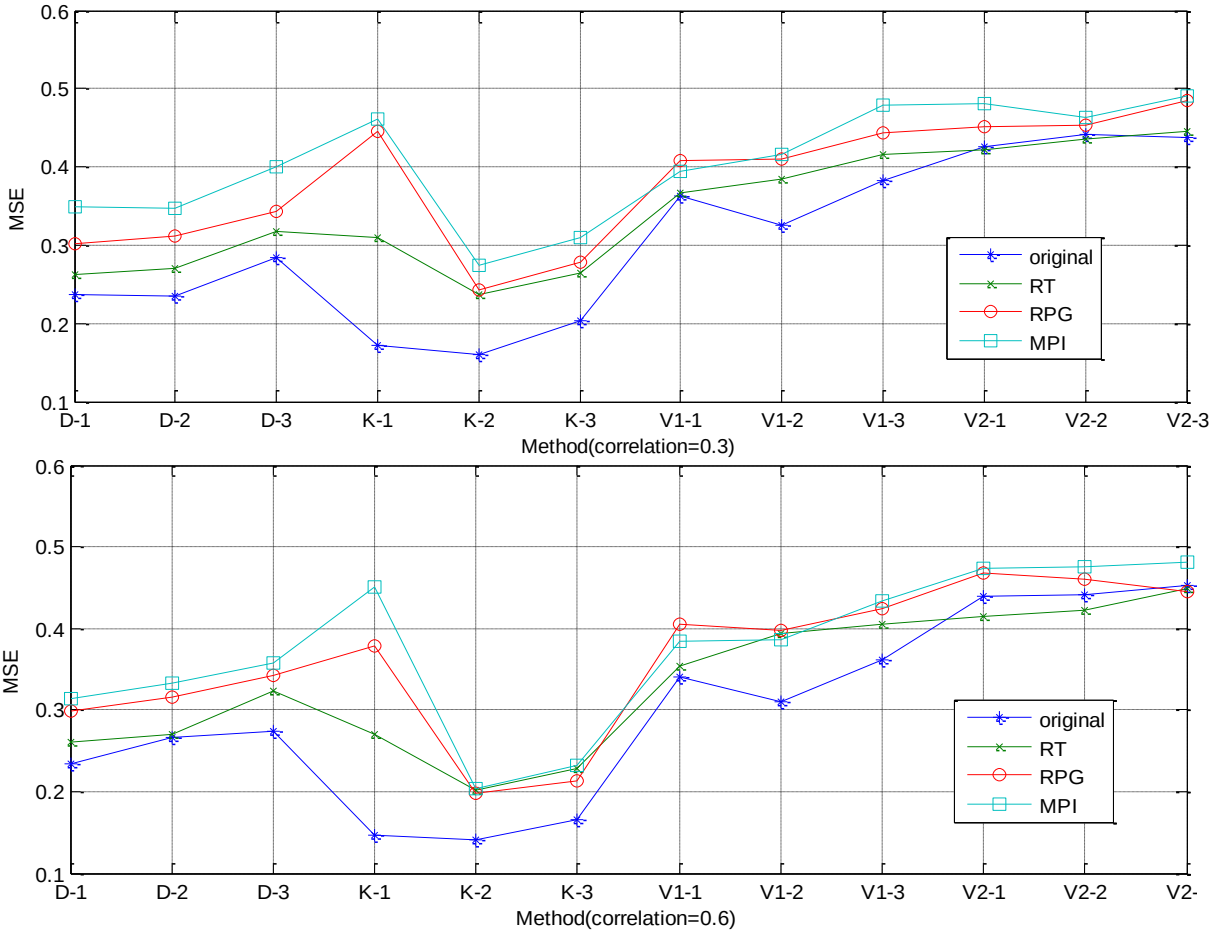
Figure 1 shows that the differences in bias between two arbitrary dimensions of each method were negligible regardless of item selection and exposure control methods. Moreover, one can observe from Figure 1 that the bias associated with D-optimality, V1, and V2 were similar, while greater than the bias produced by KLP which indicates that KLP outperformed other item selection method and effect of item exposure controlling methods on KLP and other ability estimation methods were negligible small.

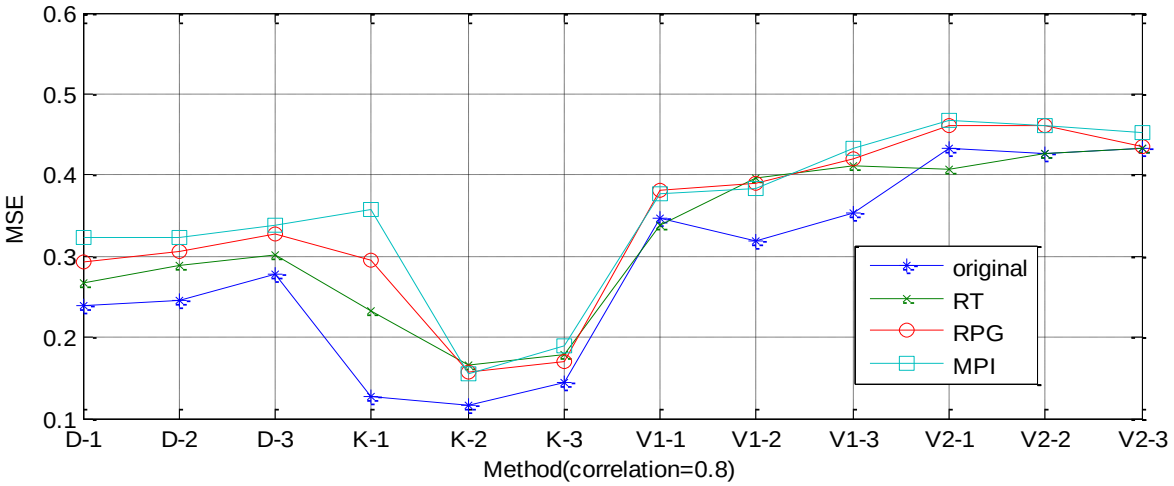
Figure 2 presents the distribution of the MSEs of each ability dimension across the different item selection and exposure controlling methods at each correlation level.

MSE statistics provided in Figure 2 shows that, for each dimension, KLP produces the smallest MSE and it was followed by D-optimality, V1, and V2. Generally, it is easy to sort the item selection methods into descending order of KLP, D-optimality, V1, and V2 according to their measurement precision. All three item exposure strategies led to an increase in MSE except for V2 item selection method. The MSE of V2 was larger than that of V2-RT in most of the cases. The decreased measurement precision may result from the characteristics of V2 in improving the item bank utility. Overall, measurement precision tends to decrease when an exposure controlling method is employed

The effects of item exposure control methods on the psychometric precision were checked through three aspects. First, from Figure 1, the item exposure strategies had no significant effect on the bias, since the biases produced by the same item selection methods using different exposure control methods were similar. Furthermore, when the item exposure control methods were combined with D-optimality, KLP, or V2, their performance differed considerably in terms of the measurement precision. However, all the item exposure control methods yielded similar measurement precision when combined with V1. In addition, a higher level of ability correlation seems to narrow the gap in the precision generated by different exposure control methods when combined with the same item selection method.

Finally, the RT exposure controlling method always produced the lowest MSE values, thus, giving higher measurement precision compared to RPG and MPI. Although their precision under different item selection indices varied to some degree, RPG and MPI performed similarly. The performance of RT and RPG was in accordance with that reported by Wang et al. (2011). Overall, the general order of different exposure control methods sorted by decreasing measurement precision was RT, RPG, and MPI, respectively.





(Note: Original=items selection methods without item exposure controlling strategies; D=D-optimality; K=KLP; '- 1','-2', and '-3'denote the first, second and third dimensions)

Figure 2. MSE of Each Ability Dimension Under Different Item Selection and Exposure Controlling Methods

Results of Item Exposure Rates

The item exposure rates and chi-square statistics associated with each item selection method with and without exposure controlling were presented in Table 1 and distribution of these statistics across different conditions were depicted in Figure 3 and Figure 4, respectively.

First, it is easy to infer from Table 1 that the exposure rates were distributed unevenly for D-optimality, KLP, V1, and V2. For instance, D-optimality and KLP yielded the largest test overlap and overexposed item rates and the lowest item bank usage rates which were depicted in Figure 3. Although the number of never-reached items in V1 and V2 was close to 0, and the test overlap rates and χ^2 values were smaller than those of D-optimality and KLP, yet, these exposure rate control methods still produced unsatisfactory item exposure rate distribution. These characteristics can be clearly observed in Figure 4(a), where the exposure rates are depicted in ascending order for each of the four item selection indices. In addition, the results for V1 and V2 obtained from this study coincide with those reported by Yao (2014a).

Table 1. Item Exposure Statistics Associated with Each Method

Item selection method	Exposure controlling method	*Overlap rate*			χ^2		
		r=.30	*r*=.60	*r*=0.80	*r*=.30	*r*=.60	*r*=0.80
D-Optimality	without exposure controlling	0.408	0.23	0.23	152.6	75.14	75.14
	RPG	0.067	0.065	0.068	3.78	2.53	3.97
	RT	0.123	0.122	0.123	25.63	24.89	24.86
	MPI	0.075	0.073	0.069	0.97	0.974	0.96
KLP	without exposure controlling	0.145	0.238	0.325	42.02	78.54	96.15
	RPG	0.078	0.074	0.074	7.23	3.40	3.45
	RT	0.121	0.119	0.118	24.45	23.47	23.10
	MPI	0.087	0.098	0.098	10.35	14.29	14.19
V1	without exposure controlling	0.253	0.241	0.237	83.5	78.78	76.29
	RPG	0.124	0.124	0.124	25.90	25.95	25.83
	RT	0.099	0.101	0.098	14.76	14.72	14.84
	MPI	0.072	0.073	0.072	2.52	2.59	2.55
V2	without exposure controlling	0.114	0.113	0.113	21.37	20.83	20.81
	RPG	0.124	0.125	0.124	15.89	25.92	15.90
	RT	0.092	0.086	0.093	11.64	8.61	11.88
	MPI	0.074	0.077	0.074	3.29	4.44	3.29

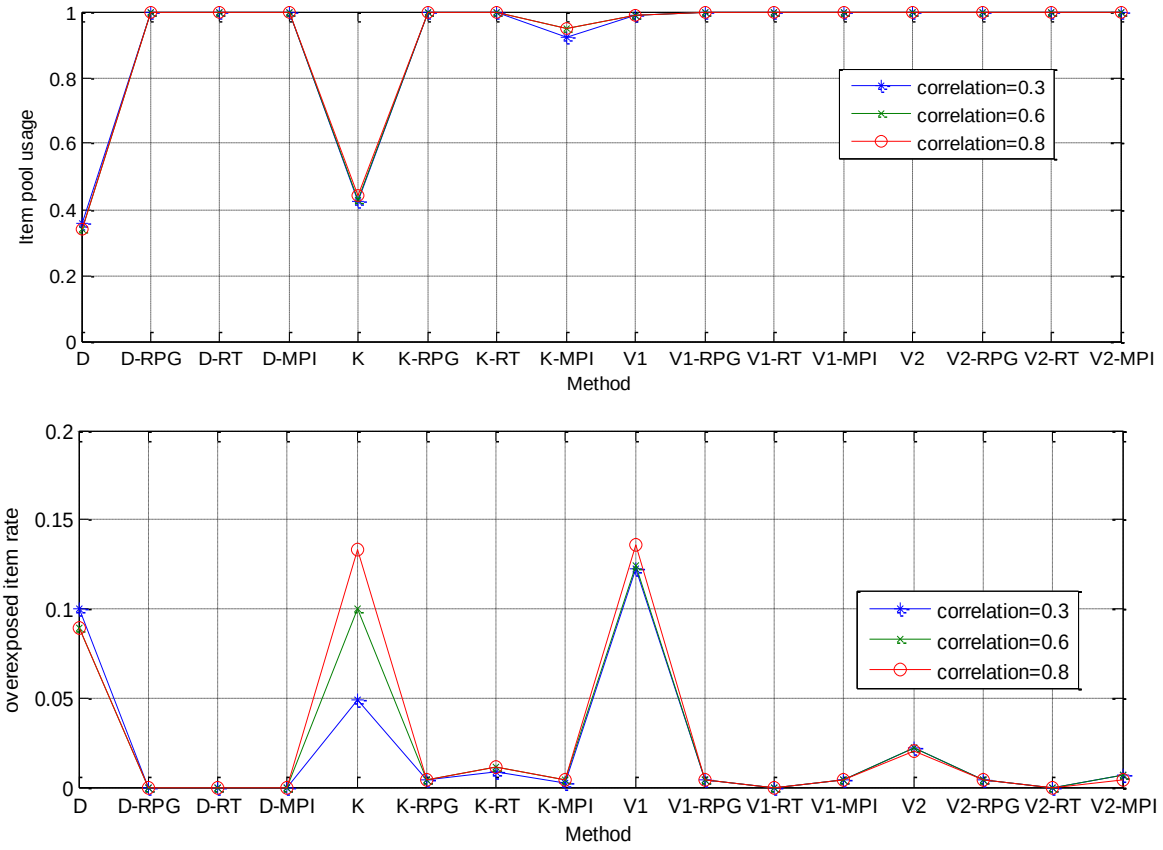


Figure 3. Item Bank Usage and Overexposed Item Rates for Each Method Under Different Correlations.

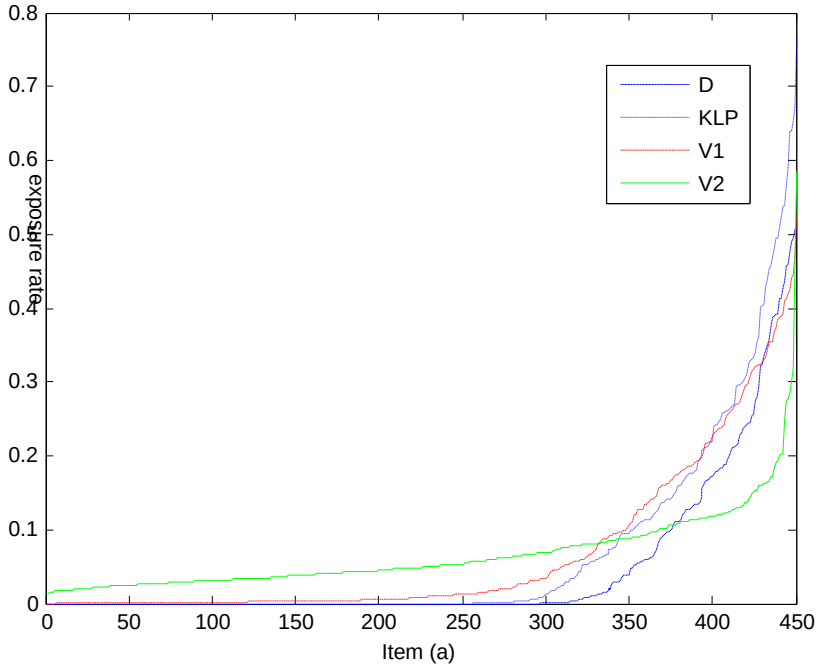
Second, all the exposure control methods improved the uniformity of exposure rates substantially in terms of increasing item bank usage and decreasing the overexposed item rates, test overlap rates, and χ^2 statistics. Although MPI performed similarly, RPG outperformed the other methods in most cases. It is apparent that all the item exposure distributions followed the same pattern when different item selection indices were combined with the same exposure control method. Hence, Figure 4(b) only illustrates the exposure rate distributions of the exposure control strategies combined with KLP.

In addition, different characteristics of the item exposure rate distribution were observed in different item exposure control methods. One can observe from Figure 3 that the item bank usage rate reaches 100% for all methods except KLP-MPI condition. In other words, all item exposure methods improve the item bank usage substantially. Checking the overexposed items, both RPG and MPI produced more overexposed items than RT under most test conditions. Generally, RT was able to control the item exposure rates to be lower than the allowable maximum value, whereas both RPG and MPI resulted in some items with exposure rates greater than 0.2.

Further, it is worth pointing out some special findings when it comes to discussing certain exposure control methods. First, compared to D-MPI, V1-MPI, and V2-MPI, KLP-MPI generated a more unbalanced item exposure rate distribution. Second, when RPG was used with V1 or V2, there were always one or two items exposed to everyone taking the test. The internal results of V1-RPG and V2-RPG revealed that many error variance values in Matlab were labeled "NaN" in the case of choosing the first or second item. In other words, it can be inferred that the overexposed items in V1-RPG and V2-RPG were mainly due to the non-distinctive item information matrix in V1 and V2.

Furthermore, the test overlap rate and χ^2 of V1-RPG and V2-RPG were affected by the first one or two administered items accordingly.

4(a) the four item selection indices without item exposure control



4(b) the three item exposure control methods combined with KLP.

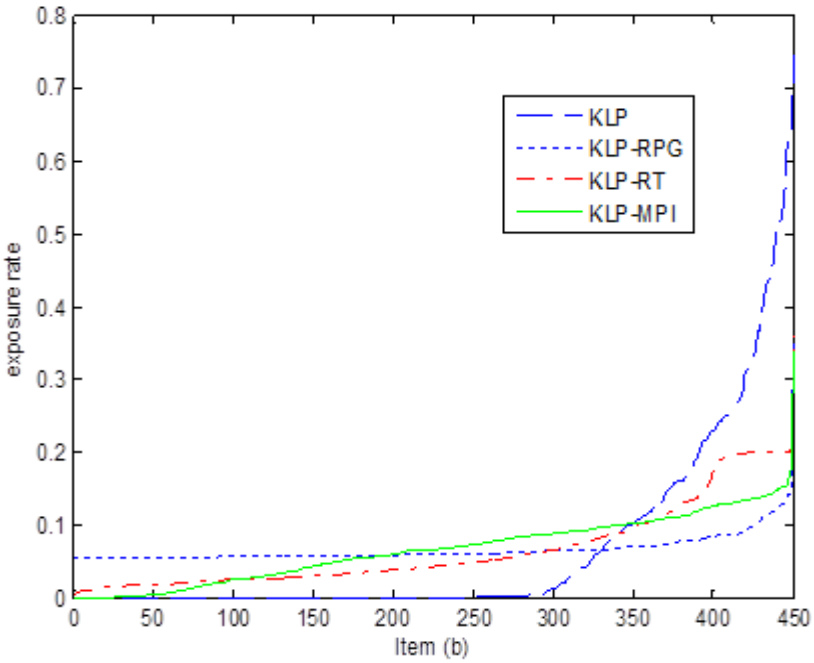


Figure 4. Item Exposure Rates of Different Methods Under the Correlation of 0.6

Overall, although the item exposure control strategies produced different patterns of item exposure rates, they all considerably improved the balance of the item exposure distribution. This can be seen from comparing Figure 4(a) and 4(b). In addition, one can infer from the results that there appear to be trade-off between the measurement precision and employing the item exposure controlling methods.

CONCLUSIONS AND DISCUSSIONS

Many studies have acknowledged the advantages of CAT over P&P tests and computer-based tests with respect to the decrease in test length, increase in measurement precision, and better model fits. Along with the obvious advantages of MCAT, choosing the most appropriate item selection rule is a vital step for a successful application (Wang & Chang, 2011). Although the proposed item selection methods yield good results in precision, they are vulnerable to the issue of dealing with overexposed items (those that are used too often) and underexposed items (used too rarely). As a solution to this problem, different item exposure control methods have been adopted and used together with different item selection methods.

This study has examined the performance of four item selection methods combined with different exposure control methods in MCAT. Simulations showed that V2 outperformed D-optimality, KLP, and V1 with respect to higher item bank usage rates, fewer overexposed items, and lower test overlap rates. Generally, the results of all item selection methods without using item exposure control were unsatisfactory with respect to item exposure statistics. The results also indicate that without using item exposure control, the item selection indices could be sorted in order of psychometric precision as KLP, D-optimality, V1, and V2. In addition, when using item exposure control methods, the measurement precision tended to decrease for all item selection method.

When the item exposure rate distribution obtained from different item exposure control methods were compared, the RPG and MPI outperformed the other methods in most cases, while the RT method showed the worst performance. Furthermore, each item exposure control method yielded the same exposure rate pattern under different item selection methods. When it comes to comparing the measurement precision, the performance of the different exposure control methods could be ordered as RT, RPG, and MPI. This kind of trade-off between measurement precision, utility of item bank, and evenness of item exposure rate has been observed in many studies (Chang & Twu, 1998). In other words, the measurement precision needs to be sacrificed, to some extent, to keep the exposure rate at the desired value.

Both the present study and the work of Wang et al. (2011) showed that the measurement precision of the RT method was higher than that of the RPG method under the same test conditions, and the RT method performed slightly worse than RPG in the evenness of the item exposure distribution. In conclusion, among the three exposure control methods examined in this study, both RT and RPG offer balanced precision and item exposure control, whereas MPI performed well in controlling the item exposure rate with a noticeable loss in precision.

Several issues regarding item selection methods for MCAT deserve further investigation. First, although D-optimality, V1, and V2 are much faster than KLP, the run-time usually increases with the number of test dimensions. As a consequence, time-consuming methods can hinder the practice of MCAT in dealing with complex test conditions. In fact, the benefits of MCAT over unidimensional CAT mainly lie in the detailed cognitive information obtained based on multiple dimensions. Hence, there is a need for more work on algorithms that reduce the computation time of the item selection methods, or simplified and valid item selection methods based on existing rules, such as the two simplified KL indexes provided by Wang et al. (2011).

Second, the test measurement precision of each dimension can be guaranteed by most MCAT item selection methods automatically, but thousands of other constraints are encountered in real tests. Hence, it would be useful to examine how to deal with non-statistical constraints in MCAT.

Third, polytomous items such as essay-type and constructed-response items have now begun to appear in CAT (Bejar, 1991). There is no doubt that research on polytomous items will increase in popularity. However, most current research on MCAT deals with dichotomous items. Thus, it is important for researchers to propose item selection methods or extend methods for dichotomous items, such as the mutual information index, KL, and Shannon entropy, to deal with polytomous items.

REFERENCES

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. *Journal of Applied Psychology, 76(4),* doi:522-532. 10.1037/0021-9010.76.4.522

Bloxom, B. M., & Vale, C. D. (1987, June). *Multidimensional adaptive testing: A procedure for sequential estimation of the posterior centriod and dispersion of theta*. Paper presented at the meeting of the Psychometric Society, Montreal, Canada.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement, 27*, 395-414. doi: 10.1177/0146621603258350

Chang, S. W., & Twu, B. Y. (September 1998). *A comparative study of item exposure control methods in computerized adaptive testing*. ACT Research Report Series, ACT-RR-98-3.

Chang, H.: H., & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222. doi: 10.1177/01466210122032181

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145. doi:10.1111/j.1745-3984.2003.tb01100.x

Cheng, Y., & Chang, H. H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British journal of mathematical and statistical psychology, 62*, 369–383. doi:10.1348/000711008X304376

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement, 46*, 84-103. doi:0.1111/j.1745-3984.2009.01070.x

Huebner, A. R., Wang, C., Quinlan, K., & Seubert, L. (2016). Item exposure control for multidimensional computer adaptive testing under maximum likelihood and expected a posteriori estimation. *Behavior Research Methods, 48*(4), 1443-1453. doi:10.3758/s13428-015-0659-z

Lee, Y. H., Ip, E. H., & Fuh, C. D. (2008). A strategy for controlling item exposure in multidimensional computerized adaptive testing. *Educational and Psychological Measurement, 68*, 215-232. doi:10.1177/0013164407307007

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20*, 389-404. doi:10.1177/014662169602000406

McKinley, R. L., & Reckase, M. D. (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82-1). American College Testing, Iowa City, IA. http://www.dtic.mil/dtic/tr/fulltext/u2/a125099.pdf

Mulder, J., & van der Linden, W. J. (2009). Multidimensional adaptive testing with optimal design criteria. *Psychometrika, 74*, 273-296. doi: 10.1007/s11336-008-9097-5

Mulder, J., & van der Linden, W. J. (2010). Multidimensional adaptive testing with Kullback-Leibler information item selection. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing, statistics for social and behavioral sciences(77-101)*. Springer, New York.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-354. doi: 10.1007/BF02294343

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 94–5). Princeton, NJ: Educational Testing Service. doi:10.1002/j.2333-8504.1994.tb01578.x. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1994.tb01578.x

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–75.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association (pp. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (1999). Multidimensional adaptive testing with a minimum error-variance criterion. *Journal of Educational and Behavioral Statistics, 24(4)*, 398-412. doi:10.3102/10769986024004398

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32(4)*, 398-418. doi: 10.3102/1076998606298044

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67(4)*, 575-588. doi:10.1007/BF02295132

Wang, C., & Chang, H. H. (2011). Item selection in multidimensional computerized adaptive testing-gaining information from different angles. *Psychometrika, 76(3)*, 363-384. DOI: 10.1007/s11336-011-9215-7

Wang, C., Chang, H. H., & Boughton, K. A. (2011). Kullback-Leibler information and its applications in multidimensional adaptive testing. *Psychometrika, 76(1)*, 13-39. DOI: 10.1007/s11336-010-9186-0

Wang, C., Chang, H. H., & Boughton, K. A. (2013). Deriving stopping rules for multidimensional computerized adaptive testing. *Applied Psychological Measurement, 37*(2), 99-122. DOI: 10.1177/0146621612463422

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement, 48(3)*, 255-273. DOI: 10.1111/j.1745-3984.2011.00145.x

Wang, W. C., & Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychologica Measurement, 28(5)*, 295–316. DOI: 10.1177/0146621604265938.

Yao, L. (2010). Reporting valid and reliability overall score and domain scores. *Journal of Educational Measurement, 47(3)*, 339-360. doi:10.1111/j.1745-3984.2010.00117.x

Yao, L. (2012). Multidimensional CAT item selection methods for domain scores and composite scores: Theory and applications. *Psychometrika, 77 (3)*, 495-523. doi: 10.1007/s11336-012-9265-5.

Yao, L. (2014a). Multidimensional CAT item selection methods for domain scores and composite scores with item exposure control and content constrains. *Journal of Educational Measurement, 51(1)*, 18-38. doi:10.1111/jedm.12032

Yao, L. (2014b). Multidimensional item response theory for score reporting. In Cheng, Y. & Chang, H.-H. (Eds.), *Advancing methodologies to support both summative and formative assessments (147-182)*. Charlotte, NC: Information Age.

Yao, L., Pommerich, M., & Segall, D. O. (2014). Using multidimensional CAT to administer a short, yet precise, screening test. *Applied Psychological Measurement, 38(8)*, 614-631. doi:10.1177/0146621614541514

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30(6)*, 3-23. doi:10.1177/0146621605284537

Çok Boyutlu Bilgisayar Ortamında Bireyselleştirilmiş Testlerde Madde Kullanım-Sıklığı Yöntemlerinin Madde Seçim Yöntemleri Üzerindeki Etkisinin Incelenmesi

Giriş

Binlerce öğrencinin aynı oturumda aynı sorulara cevap verdiği geleneksel test yöntemine alternatif olarak, öğrencilerin yetenek düzeyleri ile madde özelliklerinin bilgisayar ortamında eşleştirildiği bilgisayar ortamında bireyselleştirilmiş test yöntemleri her geçen gün yaygınlaşmaktadır. Bireyselleştirilmiş test uygulamalarının yaygınlaşmasında, geleneksel kâğıt kalem testlerine göre, uygulanmasının daha az zaman alması, testteki madde sayısını önemli ölçüde azaltması ve test biter bitmez bireye dönüt verebilmesi gibi faktörlerin etkili olduğu söylenebilir. Bireyselleştirilmiş

testlerin bir diğer avantajı ise tek boyutlu, çok boyutlu madde tepki kuramları (MTK) veya bilişsel tanı modelleri gibi farklı ölçme modellerinin (measurement models) kullanılmasına olanak sağlamasıdır. Farklı ölçme modellerinin kullanılmasına olanak sağlaması hem model-veri uyumunun incelenmesi hem de farklı puanlama yöntemlerinin kullanılmasına olanak sağlaması acısından önemli görülmektedir.

Çok boyutlu bilgisayar ortamında bireyselleştirilmiş testler ise hem çok boyutlu MTK modellerinin kullanılmasına olanak sağlaması hem de bireyselleştirilmiş olması acısından diğer yöntemlere göre avantajlı görülmektedir. Diğer taraftan farklı madde ve test seçme algoritmalarının kullanıldığı bireysel testlere ilişkin yapılan birçok çalışmada, çok boyutlu bireyselleştirilmiş testlerin tek boyutlu bireyselleştirilmiş testlere göre daha avantajlı olduğunu vurgulamaktadır. Örneğin, Segall (1996) gerçek verilere dayalı yapmış olduğu simülasyon çalışmasında tek boyutlu bireyselleştirilmiş test uygulamaları ile karşılaştırıldığında, çok boyutlu bireyselleştirilmiş testlerin test uzunluğunun üçte-bir oranında daha az olduğu ve benzer veya daha yüksek güvenirlik katsayılarına sahip olduğu bulgusuna ulaşmıştır. Luecht (1996) Yapmış olduğu çalışmada çok boyutlu bireyselleştirilmiş testlerin test uzunluğunu %25 ile %40 oranında azalttığını belirtmiştir. Ayrıca çok boyutlu modeller öğrencinin birden fazla yeteneğinin aynı anda ölçülmesine olanak sağladığından bireyin ölçülen yeteneği hakkında daha fazla bilgi sağlamaktadır. Bundan dolayı bazı geniş ölçekli test uygulamalarında tek boyutlu bireyselleştirilmiş test yerine çok boyutlu bireyselleştirilmiş testler kullanılmaktadır. Nitekim Terra Nova (Yao, 2010), American College Testing (ACT) (Veldkamp & van der Linden, 2002) ve ASVAB (Segall, 1996; Yao, 2012, 2014a) gibi testlerde gerçek madde havuzları kullanılarak çok boyutlu bireyselleştirilmiş test yöntemleri kullanılmıştır.

Çok boyutlu bireyselleştirilmiş test uygulamalarında güvenilir ve geçerli sonuçlar elde edilebilmesi ve başarılı bir şekilde uygulanabilmesinde madde secim yöntemleri önemli bir yere sahiptir (Wang & Chang, 2011). Fakat güvenilir ve geçerli sonuçlar vermelerine karşın bazı maddelerin sık uygulanması (overexposed items) veya az uygulanması (underexposed items) problemlerini çözmede yetersiz kalmaktadırlar. Bu probleme bir çözüm olarak farklı madde kullanım sıklığı yöntemleri geliştirilip, madde seçim yöntemleri ile birlikte uygulanmaya başlanmıştır.

Bu araştırmada çok boyutlu bireyselleştirilmiş testlerde kullanılan farklı madde kullanım sıklığı kontrol yöntemlerinin madde seçim yöntemleri üzerindeki etkisinin incelenmesi amaçlanmaktadır. Ayrıca, bu çalışmada madde kullanım sıklığı kontrol yöntemlerinden restrictive threshold (RT) ve restrictive progressive (RPG) yöntemlerinin madde kullanım sıklığı oranını ve diğer maddelere göre daha az uygulanan maddelerin kullanım sıklığını nasıl etkilediği incelenmiştir.

Yöntem

Bu çalışmada Monte Carlo simülasyon yöntemi ile dört farklı madde secim yönteminin farklı madde kullanım sıklığı yöntemlerinin kullanıldığı ve kullanılmadığı durumlardaki performansları karşılaştırılmıştır. Çok boyutlu MTK ya dayalı modellerin kullanıldığı simülasyon çalışmalarında genellikle boyut olarak iki veya üç boyut, madde ve yetenek parametresini kestirmek için ise çok boyutlu modellerden ise 2 parametreli veya 3 parametreli MTK modelleri tercih edildiği görülmektedir. (van der Linden, 1999; Veldkamp & van der Linden, 2002; Lee et al., 2008; Mulder & van der Linden, 2009; Finkelman et al., 2009; Wang, Chang, & Boughton, 2013; Wang & Chang, 2011). Bu simülasyon çalışmasında madde ve yetenek parametrelerinin simülasyonunda 2-parametreli MTK modelleri kullanılmış ve testler üç boyuttan oluşmaktadır. Özellikle madde havuzunda yer alan 450 maddeye ait ayırt edicilik parametreleri (a_{j1}, a_{j2}, a_{j3}) logaritmik normal dağılımdan üretilirken ($\log N(0, 0.5)$) madde güçlük parametreleri ise standart normal dağılımdan ($N(0,1)$) üretilmiştir. Her bir test için örneklem büyüklüğü 5000 olarak belirlenmiş ve bireylerin maddelere verdiği cevaplar çok değişkenli normal dağılımdan üretilmiştir. Nitekim daha önceki çalışmalarda benzer simülasyon koşulları kullanılmıştır (Wang & Chang, 2011; Yao, Pommerich, & Segall, 2014; Wang et al., 2013).

Bu çalışmada, madde secim yöntemlerinden, D-optimality, Kullback–Leibler bilgi yontemi (Kullback–Leibler information-KLP), V1 (the minimized error variance of linear combination score with equal weight) ve V2 (the composite score with optimized weight) yöntemleri kullanılmıştır. Ayrıca, madde kullanım sıklığını kontrol etmek amacıyla tek boyutlu bireyselleştirilmiş testler için geliştirilen MPI (the maximum priority index) ve bilişsel tanı modelleri için geliştirilen RT ve RPG yöntemleri kullanılmıştır. Test sürecinde yetenek parametrelerinin kestirilmesi ve güncellenmesi için Bayesyen yetenek kestirim yöntemlerinden MAP (maximum a posteriori) yöntemi kullanılmıştır. Belirlenen her bir koşul için 100 tekrar yapılmıştır.

Yukarıda belirtilen farklı çok boyutlu bireyselleştirilirmiş test koşullarından elde edilen yetenek parametrelerini karşılaştırmak için yanlılık ve standart hata ortalamaları hesaplanmıştır. Madde kullanım sıklığı yöntemlerinin etkisini incelemek için ise her bir koşula ait (a) hiç uygulanmayan madde sayısı (b) kullanım sıklığı oranı 0,2`den yüksek madde sayısı (c) ki-kare istatistiği ve (d) çakışma oranı (test overlap) istatistikleri kullanılmıştır.

Sonuç ve Tartışma

Bu çalışmada dört farklı madde seçim yöntemi ile birlikte farklı madde kullanım sıklığı yöntemlerinin kullanıldığı çok boyutlu bireyselleştirilmiş testlerin performansları karşılaştırılarak, madde kullanım sıklığı yöntemlerinin madde seçim yöntemleri üzerindeki etkisi incelenmiştir. Araştırma sonucunda, V2 madde secim yönteminin madde havuzu kullanım oranı, sık uygulanan madde oranı ve testlerdeki madde çakışma oranı açısından diğer madde secim yöntemlerine göre daha iyi sonuç verdiği bulgusuna ulaşılmıştır. Buna karşın, genel olarak, dört madde seçim yönteminin de madde kullanım sıklığı istatistikleri açısından yetersiz olduğu söylenebilir.

Madde kullanım sıklığı oranlarının dağılımı incelendiğinde, RT madde kullanım sıklığı kontrol yöntemine göre, RPG ve MPI yöntemlerinin daha iyi sonuç verdiği görülmektedir. Diğer taraftan, madde kullanım sıklığı yöntemlerinin diğer madde seçim yöntemleri ile birlikte uygulandığında maddelerin kullanım sıklığı oranı dağılımlarının benzer olduğu bulgusuna ulaşılmıştır. Ölçmenin kesinliği (measurement precision) istatistiklerine göre karşılaştırıldığında, RT yönteminin en yüksek güvenirliğe sahip olduğu ve bunu RPG ve MPI yöntemlerinin takip ettiği görülmektedir. Bu sonuçlara göre madde havuzu kullanımı ve madde kullanım sıklığı oranlarının eşitliğinin sağlanması için madde kullanım sıklığı kontrol yöntemleri uygulandığında, ölçmenin kesinliğinde belli oranda düşüşün olacağı gerçeğinin göz önünde bulundurulması gerekir (Chang & Twu, 1998). Diğer bir değişle madde kullanım sıklığı oranını istenilen düzeyde tutmak ölçmenin kesinliğinde belirli bir düzeyde düşüşü göze almayı gerektirir.

Bu çalışmada maddelerin ikili puanlandığı (0,1) çok boyutlu bireyselleştirilmiş testlerde farklı madde kullanım sıklığı yöntemlerinin madde seçim yöntemleri üzerindeki etkisi incelenmiştir. Benzer koşulların farklı madde türlerinden oluşan (örneğin çoklu puanlanan maddeler) bireyselleştirilmiş testlerde de incelenmesi önerilmektedir. Ayrıca bu çalışma çok boyutlu bireyselleştirilmiş testlerde kullanılan madde seçim ve madde kullanım sıklığı yöntemleri ile sınırlıdır. Farklı yetenek kestirim yöntemleri ve durdurma kurallarının uygulandığı test koşullarının çok boyutlu bireyselleştirilmiş testler üzerindeki etkisinin incelenmesi önerilmektedir.