ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2015; 6(2); 313-330

Applying Item Response Theory Models to Entrance Examination for Graduate Studies: Practical Issues and Insights

Madde Tepki Kuramının Akademik Personel ve Lisansüstü Eğitimi Giriş Sınavı'na Uyarlanması: Uygulamadaki Sorunlar ve Öneriler

Okan BULUT *

Abstract
Item response theory is a psychometric framework for the design, analysis, and scaling of standardized assessments, psychological instruments, and other measurement tools. Despite its increasing use in educational and psychological assessments across many countries around the world, it has not been applied to any large-scale assessment in Turkey. The purpose of this study is to investigate the fit of unidimensional item response theory models to the Entrance Examination for Graduate Studies which is a high-stake large-scale assessment in Turkey required for applying to graduate programs in Turkish universities. Model assumptions of item response modeling, such as unidimensionality, local independence, and measurement invariance, are examined. Also, model-specific assumptions, such as equal item discrimination and minimal guessing, are evaluated. Findings of this study suggest that the three-parameter IRT model shows the best model-data fit for the Entrance Examination for Graduate Studies. Also, the results of this study highlight potential issues that need to be addressed, such as high omit rates, speededness of the test, and aberrant guessing behaviors.

Key Words: Item response theory, large-scale assessment, test development, model fit.

Öz
Madde tepki kuramı standart testler, psikolojik envanterler ve diğer ölçme aletlerinin tasarımı, analizi ve ölçeklendirilmesinde kullanılan statistiksel bir modeldir. Dünyadaki birçok ülkede madde tepki kuramının ölçme ve değerlendirme alanındaki artan uygulamalarına karşın Türkiye'de bu yöntem geniş ölçekli sınavlara henüz uygulanmamıştır. Bu çalışmanın amacı tek boyutlu madde tepki kuramı modellerinin Türkiye'deki Akademik Personel ve Lisansüstü Eğitimi Giriş (ALES) sınavına uygulanmasını göstermektir. ALES sınavı Türk üniversitelerine yapılan yüksek lisans ve doktora başvuruları ve üniversitelerdeki akademik personelin belirlenmesi gibi birçok önemli alanda kullanılmaktadır. Madde tepki kuramının tek boyutluluk ve yerel bağımsızlık gibi temel varsayımlarının yanında belirli modellere özgü eşit madde ayırt edicilik gücü ve soruların minimum ölçüde tahmini gibi ek varsayımlar da incelenmiştir. Çalışmanın sonuçları üç parametreli lojistik modelin ALES için en uygun madde Tepki kuramı modeli olduğunu göstermiştir. ALES'te sınav süresinin yetersizliği, sınava girenlerin bazı soruları yüksek oranda cevapsız olarak geçmesi ve tipik olmayan soru tahmin davranışlarına dair sorunlara dikkat çekilmiştir.

Anahtar Kelimeler: Madde tepki kuramı, geniş ölçekli test, test geliştirme, model uyumu.

INTRODUCTION

Testing in education and psychology is mainly an attempt to measure a person's knowledge, intelligence, or other characteristics in a systematic and reliable way. Standardized testing has been the most useful evaluation method for measuring latent traits such as achievement, aptitude, and cognitive abilities. Standardized tests can provide decision-makers with useful information about applicants who apply for an undergraduate program in a university, try to obtain a driver's license, or apply for a job. In many testing situations, a complex measurement framework must be employed to define the

* Assistant Professor, University of Alberta, Faculty of Education, Edmonton, AB, CANADA, e-mail: bulut@ualberta.ca.

Geliş Tarihi: 11.09.2015
Kabul Tarihi: 18.11.2015

relationship between a latent trait and item responses, and generalize beyond the single situation in which a measurement is observed.

In educational testing, understanding what it takes to construct useful measures has only been applied in psychometrics (Wright, 1997). Initial methods to construct useful measures were based on the approach of counting concrete events. According to Thorndike (1904), someone who wants to measure a simple thing, such as spelling, is hampered by the fact that there exist no units in to measure. Even though one may observe the ability by the number of words from a list spelled correctly, the inequality of the units is still a serious issue. One might observe signs of spelling ability but would not have measured spelling (Engelhard, 1991). At this point, measurement models differentiate in terms of the use of raw data. There are two popular statistical frameworks for addressing measurement problems such as test development, test score equating, and the identification of biased items: classical test theory (CTT) and item response theory (IRT) (Hambleton & Jones, 1993).

CTT, also known as true score theory, was originally the leading framework for analyzing and developing standardized tests. Since the beginning of the 1970s, IRT has more or less replaced the role that CTT had and is now the major theoretical framework used in this scientific field (Crocker & Algina, 1986; Hambleton & Rogers, 1990; Hambleton, Swaminathan, & Rogers, 1991). The major advantage of CTT is its weak theoretical assumptions, which make CTT easy to apply in many testing situations (Hambleton & Jones, 1993). However, there are two major drawbacks of CTT compared to IRT. First, all item and person statistics derived from CTT are heavily dependent on the sample of test takers and the items used on the test. That is, depending on which test items are used and who takes the test, these statistics will dramatically change from one test administration to another. Second, because CTT focuses on the test-level information, it fails to explain the relationship between items and test scores. The lack of this information poses theoretical difficulties in measurement applications, such as test development, test equating, and test of measurement invariance. Unlike CTT, IRT primarily focuses on the item-level information based on the probabilistic distribution of examinees' success, and thus overcomes the technical issues that CTT has.

Item Response Theory

Item response theory, also known as latent trait theory, is not only a modern test theory, but also the most popular one in educational and psychological testing. IRT requires two major assumptions. First, the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities. Second, the relationship between examinees' item performance and the traits underlying item performance can be described by a monotonically increasing function. In IRT, this function is called "item characteristic function" or "item characteristic curve" (ICC). Based on this function, as the level of latent trait increases, the probability of an examinee giving a correct response to an item increases as well.

An example of ICC is shown in Figure1. The horizontal axis shows the ability (latent trait) scale. The ability in IRT is symbolized by the Greek letter theta (θ). The vertical axis shows the probability of giving a correct response to the item. The difficulty parameter (b) sets the location of the curve on the horizontal axis; it shifts the curve from left to right as the item becomes more difficult. The location of b can be found by dropping a vertical line from the inflection point to the horizontal axis. The slope of the curve is called the item discrimination parameter (a). The a-parameter is found by taking the slope of the line tangent to the ICC at the b-parameter. The steeper the curve, the more discriminating the item is, and the greater its item-total correlation. As the a-parameter decreases, the curve gets flatter until there is virtually no change in the probability across the ability continuum. Items with very low a values are not appropriate for differentiating examinees with low and high abilities, just like items with very low item total correlations. The c parameter is the lower asymptote. It is the lowest point of the curve as it moves to negative infinity on the horizontal axis. The c parameter can be used to model guessing in multiple-choice items.

Model Assumptions in Unidimensional IRT

There are two major assumptions for unidimensional IRT models. These assumptions are unidimensionality and local independence. The unidimensionality assumption requires that there is a single latent trait underlying a set of items. Hambleton et al. (1991) state that this assumption cannot be strictly met because of several cognitive, personality-related, and test-taking factors, such as level of motivation, test anxiety, ability to work quickly, etc. Finding a dominant component or factor affecting test performance is required to meet this assumption.

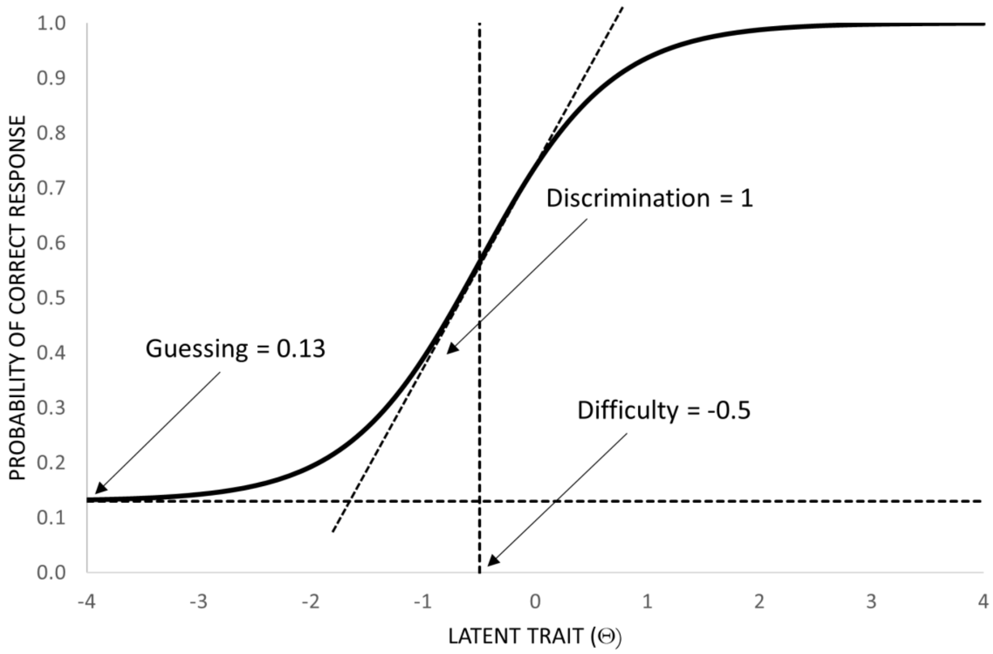

Figure 1. An Example of an Item Characteristic Curve in IRT

The local independence assumption requires the probability of a correct response by an examinee to an item not to be affected by responses given to other items in the test. In other words, after taking examinees' abilities into account, there should be no relationship between examinees' responses to different items. Therefore, high intercorrelations among the items are solely a result of the ability of the test-takers. When the trait level is controlled, local independence implies that no relationship remains between the items (Embretson & Reise, 2000).

When the assumption of unidimensionality is true, local independence is obtained. In this sense, the two concepts are equivalent (Lord, 1980). In addition to unidimensionality and local independence, there are other assumptions essential for both unidimensional and multidimensional IRT models. First, the ICC is a monotonically increasing function of the latent trait, continuous, and smooth (i.e., continuously differentiable), which results in an S-shaped curve (Hambleton et al., 1991; Raykov & Marcoulides, 2010, p. 270). Second, IRT models require invariance of item parameters and the latent trait. Item parameters are assumed to be invariant over different samples or subgroups of examinees from the population for whom the test is intended. Similarly, the latent trait needs to be invariant over different samples of test items from the population of items measuring the target ability (Baker, 1985; Hambleton et al., 1991). Third, the non-speeded test administration assumption requires that all examinees should have enough time to respond to all items in the test. The test cannot be a speeded test with binary scored items (Albanese & Forsyth, 1984). If some of these assumptions are not met, the selected IRT model is very likely not to fit to the item response data due to either poor item fit or poor person fit.

Unidimensional IRT Models

There are three frequently used IRT models for dichotomously scored test items: one-parameter, two-parameter and three-parameter IRT models. These models are the most commonly used IRT models, but there are many others – including models with the 4th parameter or upper asymptote, models that include a parameter for response time, models that include parameters for thresholds on partial credit or rating scale items, and others. The main difference between these models is the number of item parameters (a, b and c parameters as described earlier). Since the item parameters of the models are different, ICCs are also different.

The simplest IRT model is the one-parameter logistic IRT model (also known as the 1PL model). The 1PL model assumes that all of the items have the same item discrimination power and the lower asymptote (i.e., c parameters) is equal to zero for all items. The 1PL model can be shown as

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}, \qquad (1)$$

where $P_i(\theta)$ is the probability of an examinee with ability θ answering item *i* correctly, b_i is item difficulty parameter for item i, and e represents the base of the natural logarithm approximated at 2.178. The D in the equation represents a constant adjustment to the model in order to reduce the differences between the logistic IRT model and the normal ogive model to less than .01 (Crocker & Algina, 1986). The value of *D* is usually set to 1.7.

The two-parameter logistic model (also known as the 2PL model) has the same equation as the 1PL model. However, there is an additional parameter, a_i, which represents item discrimination for item *i*. It means that the item discrimination parameter varies across items, as does the item difficulty parameter. The 2PL can be shown as follows:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \qquad (2)$$

The three-parameter logistic model (also known as the 3PL model) has also a similar mathematical form. Differently from the first two models, the 3PL model includes a lower asymptote, c_i, which is the pseudo guessing level of item *i*. This additional parameter represents the probability of examinees with low ability giving a correct response to item *i* by chance. The mathematical form of the 3PL model can be written as follows:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \qquad (3)$$

The three IRT models described above can be considered as variants of each other. Among the three models, the 1PL model is the most restricted model with fixed item discrimination and zero lower asymptote for all items; whereas the 2PL model is relatively less restricted with varying item discrimination parameters and zero lower asymptote for all items. When there is random guessing (i.e., $c_i > 0$), this may result in a certain degree of inflation in the probability of correct response. This type of guessing behavior is more likely to happen among individuals with the lowest ability. The 3PL is the only model that allows for the estimation of item discrimination, item difficulty, and lower asymptote for each item.

Previous IRT Model-Fit Studies

Model-data fit studies of IRT are crucial because they provide information about the appropriateness and the adaptability of the IRT models to psychometric measures such as tests, surveys, and scales. In the literature, there are two lines of research used for investigating model-data fit in IRT. The first one is the comparison of CTT and IRT frameworks in terms of item and ability parameters by using real data and Monte Carlo simulation studies (Courville, 2005; Fan, 1998, Güler, Uyanık, & Teker, 2014; Hambleton & Jones, 1993; Progar, Sočan, & Peč, 2008). The second one is the model-data fit studies that solely focus on the application of IRT models to different instruments (e.g., tests, surveys, and scales) and provides in-depth investigations.

IRT models have been applied to various types of assessments including achievement tests, language assessments, personality inventories, and psychological instruments. One of the earliest IRT model-fit studies was conducted by Albenese and Forsyth (1984). They analyzed responses of examinees in grade 9 on five subtests of the Iowa Tests of Educational Development to compare the relative robustness of the 1PL, 2PL, and 3PL IRT models. The largest number of misfit items was observed in the 1PL model. Also, the results indicated that the modified 2PL model may provide the best representation of the data. Choi (1989) investigated the appropriateness of IRT models in language testing. The Certificate of English (FCE) from the University of Cambridge and the Test of English as a Foreign Language (TOEFL) were used for this study. The results showed that the listening subtest of the TOEFL did not meet the assumption of unidimensionality while the reading and vocabulary subtests of the FCE were convincingly unidimensional. Furthermore, the 3PL model indicated the best model-data fit for the FCE. Chernyshenko et al. (2001) compared the fit of the 2PL and 3PL IRT models to two personality assessment instruments, the US-English version of the Fifth Edition of the Sixteen Personality Factor Questionnaire and Goldberg’s 50 item Big Five Personality measure. The findings of their study suggested that the 2PL and 3PL models fit some scales reasonably well but not others. The negative keyed questions in both personality assessments led to item misfit problems across several subscales.

There have been also several IRT model-fit studies in Turkey. In an early study, Berberoglu (1990) compared the 1PL and 3PL models using the Turkish version of the Group Assessment of Logical Test (GALT) consisting of 36 multiple-choice items. The results indicated that the GALT met the assumption of having a unidimensional latent trait. In addition to the unidimensionality assumption, other IRT assumptions required for the 1PL and 3L models were also met for the GALT. In a similar study, Kilic (1999) investigated the fit of the 1PL, 2PL, and 3PL models to the four subtests of the Student Selection Test (SST) in Turkey. SST is a very high-stakes test taken by high school graduates to enter an undergraduate program in a university in Turkey. The results of this study indicated that the 3PL model fit better than the other two IRT models.

Celik (2001) also conducted a similar study using the Secondary Education Institutions Student Selection and Placement Test in Turkey. The model-fit of the three unidimensional IRT models to this test was investigated. The results indicated that the 3PL model provided a better psychometric presentation of mathematics and science subtests. Önder (2007) investigated the fit of IRT models to the data obtained from ÖZDEBİR ÖSS 2004 D-II Exam Science Test. The result of this study suggested that the most appropriate model data fit was achieved by the 3PL model, followed by the 2PL model. The most recent IRT model-fit study in Turkey was conducted by Teker, Kelecioglu, and Eroglu (2013). They investigated the fit of IRT models to the 2009 administration of Seviye Belirleme Sınavı (SBS) that is a national exam for all 8th grade students in Turkey. Their results indicated that the 3PL model was the most appropriate model for the SBS data.

The review of the IRT model-fit literature shows that the 3PL is typically the best-fitting model for multiple-choice large-scale assessments. Also, the unidimensionality assumption is more prone to be violated than the local independence assumption. In light of findings of earlier studies, this study presents an empirical investigation of IRT model-fit to address the research questions described earlier.

Purpose of the Study

To date, despite their inevitable advantages over CTT, IRT models have not been operationally used in the analysis and decision-making processes of large-scale assessments in Turkey. This study aims to demonstrate the applicability of the IRT framework using a high-stakes assessment in Turkey. The unidimensional IRT models were applied to the Entrance Examination for Graduate Studies (EEGS) that is a required examination to apply for a graduate program in Turkish universities. First, the assumptions of IRT models were examined. Then, the invariance of item and ability parameter estimates was investigated. Finally, the fit of IRT models for each subtest of the EEGS was examined to test whether the observed and theoretical distributions of IRT models overlap for the subtests of the EEGS.

METHOD

Sample

The data for this study come from the fall administration of the EEGS in 2010. The total number of examinees was 142,178. In this study, a sample of 5,000 examinees was randomly selected from the examinees that completed at least 25% of the EEGS. The descriptive statistics for the selected sample are presented in Table 1. 97.8% of the examinees completed an undergraduate program in a Turkish university and the remaining examined obtained their undergraduate degrees from a university outside of Turkey. Most of the examinees took the EEGS to apply for a graduate program or to be an academic staff in a university.

Table 1. Examinee Characteristics for the Selected Sample from the EEGS

Variable		Frequencies (f)	Percentages (%)
Gender			
	Male	2540	50.8
	Female	2460	49.2
University Location			
	Turkey	4890	97.8
	Foreign	110	2.2
Reason for taking the test			
	To apply for an academic position	872	17.4
	To apply for a graduate program	3924	78.5
	Other	204	4.1

Instrument

In this study, model-data fit analyses were carried out using the data from the 2010 administration of the EEGS. The EEGS is a large-scale assessment in Turkey that is administered twice a year by the Measurement, Selection, and Placement Center (also known as ÖSYM in Turkey). The scores from the EEGS are used for three purposes: 1) to start a graduate program in a university; 2) to determine candidates who will be sent to foreign countries for graduate education with a scholarship; and 3) to determine academic staff such as college instructors, graduate assistants, lecturers, and specialists.

The test is composed of 160 multiple-choice items with five response options. The EEGS consists of three subtests: Verbal, Quantitative 1, and Quantitative 2. The Verbal subtest includes 80 items that measure verbal reasoning abilities. The Quantitative 1 and the Quantitative 2 subtests consist of 40 items that measure mathematical and logical reasoning abilities. The Quantitative 2 subtest covers more advanced mathematical topics than the Quantitative 1 subtest.

Data Analysis

Data analysis of this study consists of three steps. The first step was preliminary data analysis. The purpose of preliminary analysis was to have an in-depth examination of the test items for any potential flaws or extreme values in the data. A CTT-based item analysis was carried out with the *psychometric* package (Fletcher, 2015) in R (R Development Core Team, 2015) for the three subtests of the EEGS. Descriptive statistics (item difficulties, point-biserial correlations, mean test scores, etc.) were obtained for the items and test scores across the subtests.

The second step was the evaluation of model assumptions. The main model assumptions, namely unidimensionality and local independence, were carefully investigated. The assumption of unidimensionality requires that the probability of successful performance by examinees on a set of items can be modeled by a mathematical model that has only one ability parameter (Dorans & Kingston, 1985). Although this is a very important assumption for IRT models, there is no simple way to assess the unidimensionality assumption. Stout's nonparametric DIMTEST (Stout, 1987), Humphreys and Montanelli's (1975) method of parallel analysis, and confirmatory factor analysis (CFA) are the most widely used methods for assessing scale unidimensionality.

In this study, the CFA approach was used to confirm the unidimensional latent structure of the Verbal, Quantitative 1, and Quantitative 2 subtests. A one-factor (i.e., unidimensional) CFA model was fit to each of the three subtests using Mplus 6 (Muthen & Muthen, 1998-2011). A robust weighted least squares (WLS) estimator with a diagonal weight matrix was used as the estimation method. The WLS estimator was selected because when dependent variables are identified as categorical, this estimator yields more accurate factor loading estimates than maximum likelihood and robust maximum likelihood estimators (Li, 2014).

Goodness-of-fit criteria, including root mean square error of approximation (RMSEA), Tucker-Lewis Index (TLI), and comparative fit index (CFI), were used to evaluate model-data fit of the one-factor CFA model for the three EEGS subtests. CFI and TLI are incremental fit indices that assess the relative improvement in fit of the selected model compared with a baseline model. Both indices range between 0.0 and 1.0 with values closer to 1.0 indicating good fit. RMSEA is an absolute fit index that is independent of sample size and thus performs well as an indicator of practical fit. For CFA models, Hu and Bentler (1999) suggested that for categorical data, RMSEA < .06, TLI > .90, and CFI > .90 indicate good fit. Based on these criteria, a satisfactory fit for the one-factor model would suggest that the test has a unidimensional structure.

There are also several methods for assessing local item dependencies in dichotomous data, such as Yen's Q_3 statistic (Yen, 1984) and the G^2 statistic (Bishop, Fienberg, & Holland, 1975; Chen & Thissen, 1997). Conditional inter-item correlations can be also used as a measure of local item independence (Ferrara, Huynh, & Baghi, 1997). In this study, to examine the assumption of local independence, inter-item correlation matrices were evaluated in a restricted range of abilities (i.e., high ability and low ability groups). For selecting low and high ability examinee groups, the 20th and 80th percentiles of total raw scores were used as the cut-off values in each subtest. The zero or close to zero off-diagonal elements of the variance-covariance or the correlation matrix for examinees within a restricted range of ability or test score scale indicate unidimensionality and that the test has met the assumption of local independence (Hambleton et al., 1991; McDonald, 1981).

To check the measurement invariance of the EEGS subtests between male and female examinees, a multi-group CFA framework (Meredith, 1993) was used. A one-factor CFA model for a dichotomous observed response, X_i, for item *i* can be written as follows:

$$X_i = \tau_i + \lambda_i \xi + \varepsilon_i, \tag{4}$$

where τ_i is the intercept for item *i*, λ_i is the factor loading for item *i*, ξ is the latent construct, and ε_i is the residual term for item *i*. To test measurement invariance across male and female examinees, a series of nested multiple group models was assessed.

Table 2 summarizes the four types of measurement invariance tests used in this study. For each test, a constrained model with fixed parameters across male and female examinees was tested against a less constrained model. The nested models were compared using a chi-square difference test as well as several model fit indices. Substantial decrease in goodness of fit between the two models indicates the violation of measurement invariance. Measurement invarance tests were conducted using the lavaan package (Rosseel, 2012) in R (R Core Team, 2015).

Table 2. Summary of Measurement Invariance Tests

Type	Constrained Parameters	Comparison Model
Configural invariance	-	-
Weak invariance	λ_i	Configural invariance
Strong invariance	τ_i, λ_i	Weak invariance
Strict invariance	$\tau_i, \lambda_i, \varepsilon_{ij}$	Strong invariance

After the unidimensionality, local independence, and measurement invariance assumptions were checked, the other model-specific assumptions were also carefully examined. The homogeneous distribution of item discrimination indices obtained from the preliminary item analysis was used to check the assumption of equal discrimination indices of the 1PL model. Performance of low-ability examinees on the most difficult questions was evaluated to check the minimal guessing assumption of the 1PL and 2PL models. The most difficult items were chosen based on the proportion-correct values obtained from the preliminary item analysis. The non-speeded test administration assumption was evaluated based on the percentages of examinees that completed the last five items of each subtest of the EEGS.

The final step of the data analysis was the comparison of model-data fit. The three subtests of the EEGS were calibrated and scored based on the 1PL, 2PL, and 3PL models, respectively. The IRT model estimation was implemented using marginal maximum likelihood estimation in Xcalibre 4.1 (Guyer & Thompson, 2011). The fit of the 1PL, 2PL, and 3PL models was compared using the Likelihood Ratio (LR) test, which is based on -2 times the difference in log-likelihoods from two nested models. The LR statistic can be computed as follows:

$$LR = -2\ln L_C - (-2\ln L_A), \qquad (5)$$

where L_C is the log likelihood of the compact model (i.e., the model with fewer item parameters) and L_A is the log likelihoods of the augmented model (i.e., the model with more item parameters).

The LR statistic is approximately distributed as chi-square (χ^2) with degrees of freedom equal to the difference in the number of parameter estimates in the two models. The significant LR statistic indicates that the augmented model fits better than the compact model. Drasgow et al. (1995) suggested that the adequacy of model fit should be also evaluated using graphical methods. In this study, in addition to the LR test for model comparison, both model fit plots at the item and test levels as well as chi-square goodness of fit statistics for individual items were used to examine the fit of IRT models to the EEGS.

RESULTS

Results of Preliminary Analysis

The results of preliminary item analysis are presented in Table 3. The results indicated that Quantitative 2 was the most difficult subtest on average among the three subtests. The average item-total correlations (i.e., point-biserial correlations) demonstrate the discriminatory level of the three subtests between high ability and low ability examinees. Based on the results in Table 3, Quantitative 2 indicated a better discriminatory power than the other two subtests. In addition, the results indicated that all of the items functioned and discriminated well. Therefore, none of the items were excluded from the subsequent analyses. The EEGS indicated high test reliability based on coefficient alpha values obtained from each subtest.

Table 3. Summary Statistics for the Items in the EEGS Subtests

Subtest	*N*	Mean Difficulty	Mean Point-Biserial	Alpha
Verbal	80	0.73	0.43	0.96
Quantitative 1	40	0.76	0.39	0.90
Quantitative 2	40	0.67	0.46	0.93

The results of preliminary analysis at the subtest level are presented in Table 4. Descriptive statistics based on total raw scores are presented for the overall sample and for each gender group separately. The scores from the three subtests had negatively skewed distributions, suggesting that most examinees in the sample obtained high scores in the EEGS. Although the minimum and maximum raw scores did not differ across gender groups, female examinees performed better than male examinees in the Verbal subtest and the male examinees outperformed the female examinees in both Quantitative 1 and 2 subtests. Especially in the Verbal subtest, the distribution of raw scores was negatively skewed as most of the students obtained high test scores.

Table 4. Summary Statistics for the Total Raw Scores from the EEGS Subtests

Subtest	Group	N	M	SD	Min	Max	Skewness	Kurtosis
Verbal	Overall	5000	58.1	15.6	20	80	-0.63	-0.68
	Male	2540	56.9	16.2	20	80	-0.54	-0.87
	Female	2460	59.3	14.8	20	80	-0.72	-0.45
Quantitative 1	Overall	5000	30.4	7.1	10	40	-0.83	-0.07
	Male	2540	31.3	6.9	10	40	-1.03	0.40
	Female	2460	29.4	7.0	10	40	-0.66	-0.36
Quantitative 2	Overall	5000	26.8	8.8	10	40	-0.42	-1.12
	Male	2540	27.7	8.9	10	40	-0.55	-0.98
	Female	2460	25.8	8.7	10	40	-0.31	-1.20

Results of Model Assumptions

The CFA results indicated that all of the three subtests had acceptable levels of model-data fit based on the model-fit criteria described earlier (see Table 5). The satisfactory model-fit for the one-factor CFA model suggested that the unidimensionality assumption of the Verbal, Quantitative 1, and Quantitative 2 subtests was adequately met. However, it should be noted that the model fit indices presented in this study may not be robust against issues such as sample size or missing item responses in the data. Therefore, the use of alternative dimensionality tests is strongly encouraged for a more detailed analysis of the unidimensionality assumption.

Table 5. Model Fit Indices for the One-Factor CFA Model

Subtest	*N* of Items	CFI	TLI	RMSEA
Verbal	80	0.90	0.89	0.01
Quantitative 1	40	0.94	0.98	0.01
Quantitative 2	40	0.97	0.99	0.01

The means of inter-item correlations of high and low ability groups were close to zero across the three subtests (see Table 6). However, the results suggested that some items in the Verbal subtest had relatively higher inter-item correlations than the items in the other two subtests. These items were mostly linked to the same reading passages on the test, which suggests that some of the items may be problematic since they depend on the same content. Therefore, although the local independence assumption was assumed to be met based on inter-item correlations, the likelihood of having locally dependent items based on the reading passages remained as a potential concern for the Verbal subtest.

Table 6. Inter-Item Correlations Obtained from the Low and High Ability Groups

Subtest	Low Ability Group				High Ability Group			
	M	*SD*	Min	Max	*M*	*SD*	Min	Max
Verbal	0.155	0.098	-0.256	0.455	0.180	0.101	-0.299	0.484
Quantitative 1	0.071	0.098	-0.114	0.267	0.074	0.101	-0.194	0.289
Quantitative 2	0.053	0.071	-0.093	0.197	0.055	0.079	-0.101	0.203

To investigate the equal item discrimination assumption for the 1PL model, the frequency distributions of item-total correlations obtained from the preliminary item analysis were analyzed graphically. As seen in Figure 2, the item-total correlations were not homogenously distributed, suggesting that the items may not have an equal discrimination power and that the assumption of equal item discrimination may not be viable for the EEGS.

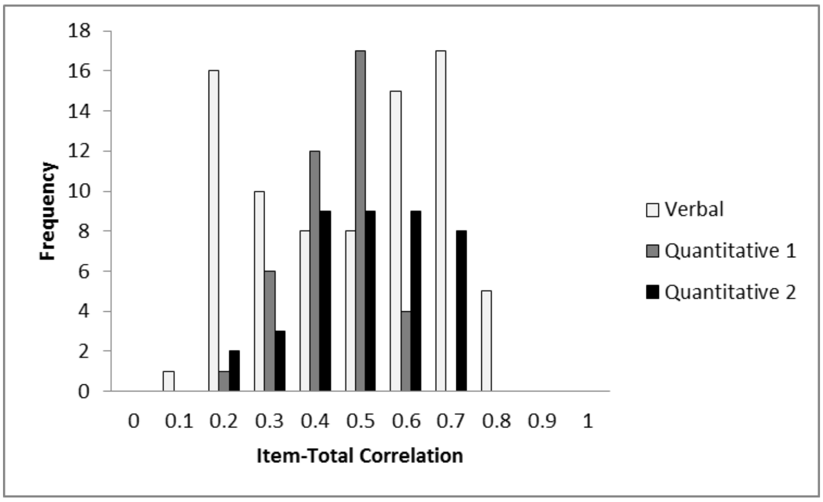

Figure 2. Frequency Distribution of Item-Total Correlations from the EEGS Subtests

The minimal guessing assumption for the 1PL and 2PL models was investigated by examining the performance of low-ability examinees (i.e., 20th percentile and below) on the most difficult items. The most difficult items were identified by selecting 10% of the items with the lowest proportion correct values. This procedure resulted in selecting eight items from the Verbal subtest, and four items from the Quantitative 1 and Quantitative 2 subtests. The results are presented in Table 7. The performance of low-ability examinees was worse than the performance of the overall sample on the difficult items. Low-ability examinees chose to skip the difficult items instead of randomly guessing. Most of the difficult items were mostly the last items on the tests and these items had high omit rates. This finding suggests that although minimal guessing assumption was met in the EEGS, high omit rates may still be a concern.

Table 7. Percentage of Correct Responses on Most Difficult Items by Low-Ability Examinees

Subtests	Items	*P*	Percent Correct	Percent Incorrect	Percent Missing
Verbal	Item 56	0.47	8.0	12.7	79.3
	Item 73	0.47	11.1	5.2	83.7
	Item 75	0.38	7.2	7.3	85.5
	Item 76	0.33	6.4	7.4	86.2
	Item 77	0.41	11.3	4.4	84.3
	Item 78	0.41	10.7	4.2	85.2
	Item 79	0.43	12.2	2.6	85.2
	Item 80	0.31	7.7	6.1	86.2
Quantitative 1	Item 30	0.44	25.6	55.6	18.8
	Item 31	0.46	15.2	26.1	58.7
	Item 32	0.38	9.4	28.8	61.8
	Item 33	0.47	16.2	17.4	66.4
Quantitative 2	Item 10	0.29	16.2	29.9	53.9
	Item 30	0.35	9.2	12.8	78.0
	Item 36	0.39	2.7	5.4	91.9
	Item 39	0.15	1.5	12.9	85.6

Note: *P* is the proportion of correct responses from the overall sample.

To check the non-speeded test administration assumption, percentages of omitted responses on the last five items in the Verbal, Quantitative 1, and Quantitative 2 subtests were examined. The results are presented in Table 8. The percentages of omitted responses on the last five items were significantly higher in the Verbal subtest than the other two subtests, regardless of difficulty levels of the items. The last five items in the Quantitative 1 subtest had low omit rates while the difficult items (item 36 and item 39) in the Quantitative 2 subtest indicated substantially higher omit rates than the other items. Those two items had high item-total correlations, suggesting that despite their high omit rates, the items discriminated high-ability and low-ability examinees very well.

Table 8. Descriptive Statistics for the Last Five Items on the EEGS Subtests

Subtest	Item	p-value	Item-total Correlation	Proportion Missing
Verbal	Item 76	0.33	0.48	0.51
	Item 77	0.41	0.50	0.53
	Item 78	0.41	0.52	0.54
	Item 79	0.43	0.51	0.54
	Item 80	0.31	0.45	0.56
Quantitative 1	Item 36	0.84	0.45	0.13
	Item 37	0.65	0.43	0.19
	Item 38	0.77	0.45	0.11
	Item 39	0.76	0.55	0.16
	Item 40	0.74	0.52	0.19
Quantitative 2	Item 36	0.39	0.61	0.53
	Item 37	0.85	0.44	0.12
	Item 38	0.80	0.49	0.18
	Item 39	0.15	0.36	0.52
	Item 40	0.66	0.61	0.29

Note: p-value is the proportion of correct responses.

One theoretical feature that makes IRT models superior over other psychometric frameworks is the invariance (i.e., equality) of item and examinee parameters from different examinee populations or measurement conditions (Rupp & Zumbo, 2006). Parameter invariance in IRT can be investigated when there are at least two examinee populations or two measurement conditions for parameter comparisons. In this study, measurement invariance of item parameters was investigated across male and female examinees using a multi-group CFA framework.

Table 9 shows the results of measurement invariance tests for the three subtests of EEGS. For all of the subtests, weak invariance was met, which suggests that the constructs indicated the same meaning across male and female examinees. In the context of IRT, fixing factor loadings across male and female examinees for testing weak invariance is analogous to fixing item discrimination parameters across male and female examinees. Weak invariance of the items in the EEGS shows that the discriminatory power of the items did not differ between male and female examinees. Strong invariance was ensured for the Quantitative 1 and Quantitative 2 subtests but not for the Verbal subtest. As explained earlier, strong invariance assumes that item intercepts are equal across groups. In the context of IRT, item intercepts are analogous to item difficulty parameters. If one group has higher or lower probability to respond to item correctly than the other group, then this affects the means of the observed item, hence affects the mean of the scale and the latent variable. In this study, significant χ^2 change in the Verbal subtest indicated non-invariance of intercepts (i.e., item difficulties) between male and female examinees. Therefore, it can be concluded that some items in the Verbal subtest were systematically easier or more difficult for one of the gender groups. Finally, strict invariance was met for none of the EEGS subtests. Strict invariance is particularly important for group comparisons based on the sum of observed item scores, because observed variance is considered as a combination of true score variance and residual variance. The violation of this invariance test suggests that the items in EEGS may not be equally reliable across male and female examinees. According to Meredith (1993), strict invariance is necessary for a fair and equitable comparison across groups. Because none of the EEGS subtests indicated strict invariance in this study, it can be concluded that test scores from the three subtests of EEGS cannot be reliably and meaningfully compared between male and female examinees.

Table 9. Results of Measurement Invariance Tests for the EEGS Subtests

Subtest	Type of Invariance Test	$\Delta\chi^2$	Δdf	CFI	RMSEA
Quantitative 1	Configural invariance	-	-	0.75	0.05
	Weak invariance	48.56	39	0.72	0.05
	Strong invariance	52.14	39	0.72	0.05
	Strict invariance	664.14*	40	0.65	0.05
Quantitative 2	Configural invariance	-	-	0.75	0.06
	Weak invariance	51.23	39	0.73	0.06
	Strong invariance	56.25	39	0.73	0.06
	Strict invariance	283.28*	40	0.71	0.07
Verbal	Configural invariance	-	-	0.68	0.04
	Weak invariance	99.56	79	0.67	0.04
	Strong invariance	204.60*	79	0.66	0.04
	Strict invariance	687.43*	80	0.54	0.05

Note: $\Delta\chi^2$ = Difference in chi-square between the two consecutive models; Δdf = Difference in degrees of freedom between the two consecutive models; * p-value < .05

Results of Model-Fit Comparison

The advantages of IRT models can be achieved only if there is a satisfactory goodness-of-fit between the model and test data (Gao, 2011). In this study, the overall fit of 1PL, 2PL, and 3PL models was compared based on the Likelihood Ratio test. Table 10 presents the results of model-fit comparisons. For all of the EEGS subtets, the 3PL indicated the best model fit, the 2PL model was the second best-fitting model, and the 1PL model indicated the worst model fit. Especially in the Verbal subtest, the difference between -2 log likelihood values of the 1PL and 2PL models was very large. Because the restricted 1PL model cannot account for the variation among the Verbal test items as much as the other two models, the resulting model fit was very poor.

Table 10. Comparison of the Three IRT Models for the EEGS Subtests

Subtest	Comparisons	
	1PL vs. 2PL	2PL vs. 3PL
Quantitative 1	1659.063 (39)*	450.756 (40)*
Quantitative 2	4608.907 (39)*	1039.093 (40)*
Verbal	18835.897(79)*	2161.296 (80)*

Note: Each cell shows the difference in -2 log likelihood values and difference in the number of estimated item parameters. * p-value < .001

Although statistical tests of goodness-of-fit are widely used in the evaluation of model-data fit, they often provide inconclusive evidence for adequate model-data fit because of their sensitivity to sample size and their insensitivity to certain forms of model-data misfit (Chernyshenko et al., 2001; Van der Wollenberg, 1982). Therefore, it is important to use graphical fit plots of ICCs and item information functions (IIFs), in addition to chi-square goodness of fit tests for single items. IIF of an item can be expressed as:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)}, \tag{6}$$

where $P_i(\theta)$ is the probability of correctly responding to item i given θ, $Q_i(\theta)$ is equal to (1-$P_i(\theta)$), and $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ given θ. In this study, IFFs from the 1PL, 2PL, and 3PL models were computed for each item in the Verbal, Quantative 1, and Quantitative 2 subtests. Misfit items were identified based on chi-square goodness of fit statistics for the items and the evaluation of ICCs (see Table 11). The results suggested that the 3PL model provided the best model-fit in the Verbal and Quantitative 2 subtests. However, in the Quantitative 1 subtest, the 2PL model indicated better model-fit based on the fewer number of misfit items. A potential reason for the worse model-fit of the 3PL model in the Quantitative 1 subtest might be omitted responses and aberrant response patterns of examinees. In the context of IRT, some examinees may have unexpected guessing behaviors which may result in high guessing parameters ($c > 0.5$) for the items. To further investigate the extent to which examinees' response patterns are consistent with expectation, person fit statistics – such as the log-likelihood person-fit statistic (Levine & Rubin, 1979) and the standardized log-likelihood statistic (Drasgow, Levine, & Williams, 1985) – can be used.

Table 11. Number of Misfit Items in the Three Subtests of the EEGS

Subtest	N	IRT Models		
		1PL	2PL	3PL
Quantitative 1	40	20 (50%)	6 (15%)	10 (25%)
Quantitative 2	40	19 (45.5%)	11 (27.5%)	7 (17.5%)
Verbal	80	34 (42.5%)	12 (15%)	11 (13.8%)

In addition to the evaluation of model-fit based on item-fit statistics and model-fit plots, marginal test reliability, test information functions, and conditional standard error of measurement (CSEM) can be useful in selecting the best-fitting IRT model. Test information function (TIF) is basically the sum of all IIFs in a given test, which provides a visual depiction of where along the trait continuum a test is most discriminating (Reise & Waller, 2002).

Figure 3 shows TIF and CSEM plots from the 1PL, 2PL, and 3PL models across the three subtests of the EEGS. The results suggested that except for the Quantitative 1 subtest, the 3PL model provided more information and less measurement error than the other two models along the ability continuum. The 2PL model was evidently better than the 1PL and 3PL models in the Quantitative 2 subtest, which also supports the findings from item misfit analyses. The performance of the 1PL model was similar to the performances of the 2PL and 3PL models only on the tails of the distributions where very low-ability and high-ability examinees were located. The marginal reliability was above .90 for all of the IRT models across the three subtests of EEGS.

DISCUSSION AND CONCLUSION

The findings of this study suggest that despite the appealing practical features of IRT, fitting IRT models to large-scale assessments requires a comprehensive investigation of model assumptions, item fit, and overall model-fit. Compared to CTT, IRT requires much stronger model assumptions, such as unidimensionality and local independence of the items. Without adequate evidence supporting the integrity of those assumptions, IRT results from an operational assessment may not be credible.

As Chernyshenko et al. (2001) pointed out; there is a strong trade-off between searching for the most appropriate IRT model that adequately describes item responses and rejecting items that do not fit a chosen model. The findings of this study suggest that more complex models (e.g., the 3PL model) tend to fit the data from large-scale assessments better than the simple models (e.g., Rasch model, 1PL model). However, it should be noted that the selection of more complex models will increase the minimum sample size required for IRT analyses, as well as limit prospective applications of IRT in practical settings (Chernyshenko et al., 2001). Although sample size may not be a concern in large-scale assessments, other possible issues in large-scale assessments, such as high omit rates and random-guessing, still remain as potential threats to the estimation of complex IRT models.

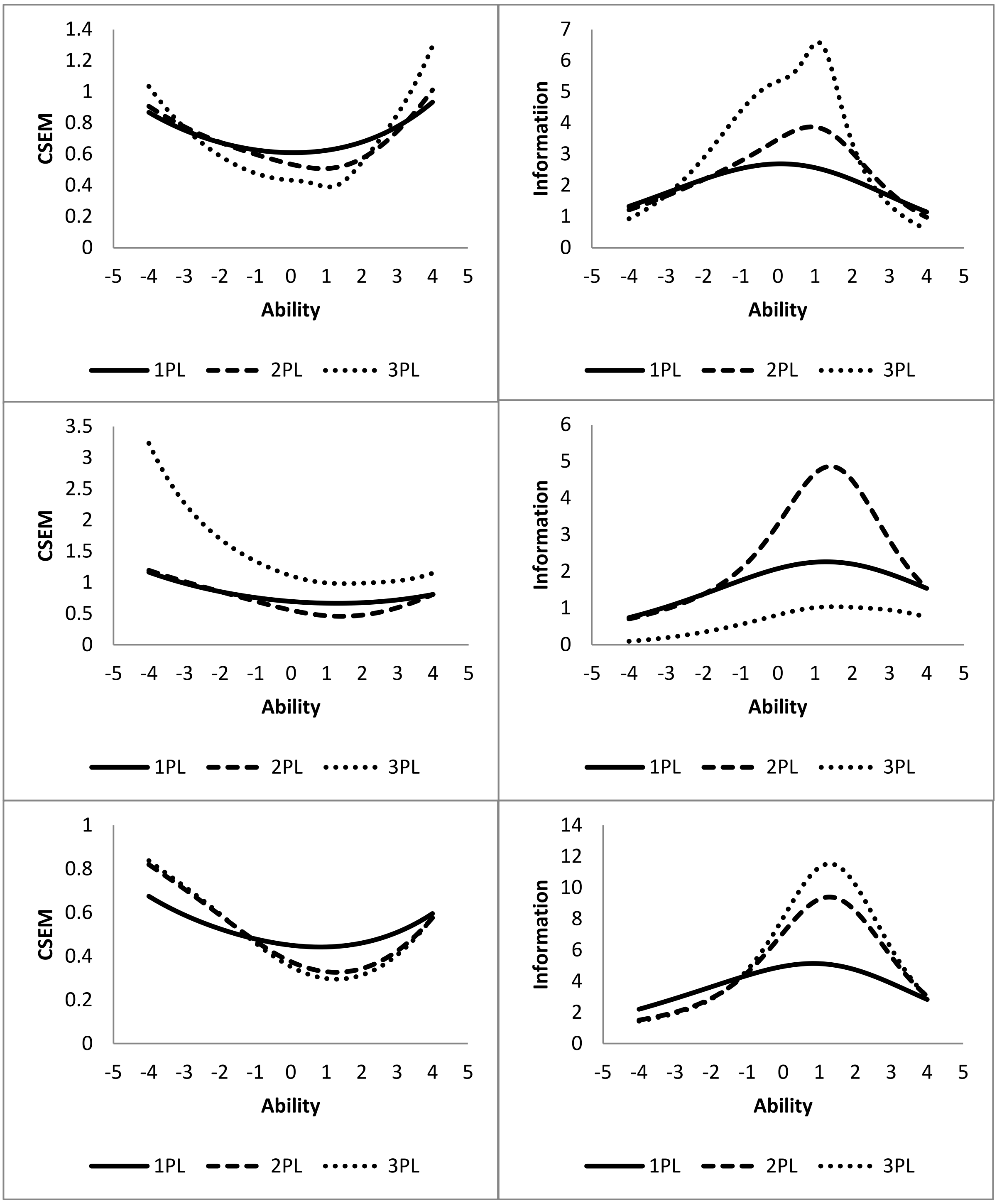

Figure 3. Conditional Standard Error of Measurement and Test Information Functions of the IRT Models for the Quantitative 1 (top), and Quantitative 2 (middle), and Verbal (bottom) Subtests

In this study, the estimation of guessing parameter was particularly problematic in the Quantitative 1 subtest because of higher difficulty levels of the items and higher omitted response rates. The assumption for the guessing parameter that every examinee has the same probability to guess an item correctly may not reflect the real guessing situation (De Ayala, 2008). Therefore, it is difficult to find the reasons of the guessing problem in the EEGS without further investigation. Also, despite the fact that the 2PL model does not account for guessing in the items, it provided better model-fit than the 3PL model in the Quantitative 1 subtest. It is because the 3PL model can produce the most accurate item parameter and ability estimates when a moderate amount of guessing is assumed (Pelton, 2002). Guessing degrades the fit of the IRT models, because the empirical ICC cannot be expected to approach zero when the theta value is small (Progar, Sočan, & Peč, 2008).

The invariance property of IRT item and ability parameters has important implications for the application of IRT to large-scale assessments. First, assuming that item parameters and ability parameters are invariant regardless of who takes the test and which items are used, computerized adaptive testing (CAT), where each examinee responds to a different set of items from the precalibrated item bank, can be implemented. For instance, Bulut and Kan (2012) demonstrated the applicability of CAT in the EEGS. Their results suggested that CAT provided highly accurate ability estimates using much fewer test items than the paper-pencil form of the EEGS. Second, the invariance property of IRT models facilitates creating comparable scores on different forms of an assessment. A linear transformation of ability estimates can equate the test scores from groups of examinees with different abilities, such as students in different grades, or, from groups of examinees with similar abilities who take the test at different times. This feature would allow producing valid and compareable scores from the EEGS across multiple test administrations. Because test scores can be placed on a common scale through equating and linking procedures, the scores can be directly compared between the examinees who might take the test at different administrations of the EEGS.

In light of the findings of the present study, future studies should focus on the impact of omitted item responses on the validity and reliability of IRT-based test scores obtained from large-scale assessments. Furthermore, more comprehensive studies are needed for understanding the invariance of the item parameters between male and female examinees. Testing differential item functioning of the EEGS items can be helpful for understanding why male and female examinees differ on the Verbal, Quantitative 1, and Quantitative 2 subtests.

REFERENCES

Albenese, M. A., & Forsyth, R. A. (1984). The one-, two-, and modified two parameter latent trait models: An empirical study of relative fit. *Educational and Psychological Measurement, 44*(2), 229–246.

Baker, F. B. (1985). *The basic of item response theory*. Portsmouth, NH: Heinemann.

Berberoglu, G. (1990). Do the Rasch and three-parameter models produce similar results in test analyses? *Journal of Human Sciences, 10*, 7–16.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to Entrance Examination for Graduate Studies in Turkey. *Eurasian Journal of Educational Research*, 49, 61–80.

Celik, D. (2001). *The fit of the one-, two- and three-parameter models of item response theory (IRT) to the ministry of national education secondary education institutions student selection and placement test data.* Unpublished master's thesis, Middle East Technical University, Ankara, Turkey.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265–289.

Chernyshenko, O. S., Stark, S., Chan, K. Y.,, Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: issues and insights. *Multivariate Behavioral Research, 36*(4), 523–562.

Choi, I. (1989). *An application of item response theory to language testing: Model-data fit studies* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Courville, T. G. (2005). *An empirical comparison of item response theory and classical test theory item/person statistics* (Unpublished doctoral dissertation). Texas A&M University.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston, Inc.

De Ayala, R. J. (2008). *The theory and practice of item response theory.* New York, NY: Guilfords Publications.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*(4), 249–262.

Drasgow, F., Levine M. V., Tsien, S., Williams B. A., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, *19*, 143–165.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67–86.

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

Engelhard, G. (1991). Thorndike, Thurstone and Rasch: A comparison of their approaches to item-invariant measurement. *Journal of Research and Development in Education, 24*(2), 45–60.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person parameters. *Educational and Psychological Measurement, 58*, 357–381.

Ferrara, S., Huynh, H., & Bagli, H. (1997). Contextual characteristics of locally dependent open-ended item clusters on a large-scale performance assessment. *Applied Measurement in Education*, *12*, 123–144.

Fletcher, T. D. (2015). psychometric: applied psychometric theory. [Computer software]. Available from http://CRAN.R-project.org/package=psychometric.

Gao, S. (2011). The exploration of the relationship between guessing and latent ability in IRT models. *Dissertations.* Paper 423.

Güler, N., Uyanık, G. K., & Teker, G. T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education, 2*, 1–6.

Guyer, R., & Thompson, N.A., (2011). *User's Manual for Xcalibre 4.1.* St. Paul MN: Assessment Systems Corporation.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R. K., & Rogers, J. H. (1990). Using item response models in educational assessments. In W. Schreiber, & K. Ingenkamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). England: NFER-Nelson.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. New York: Sage publications.

Hu, L.T., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Humphreys, L. G., & Montanelli, R. G. (1975). An investigation of the parallel analysis criterion for determining the number of common factors. *Multivariate Behavioral Research, 10*, 193–205.

Kilic, I. (1999). *The fit of one, two and three parameter models of item response theory to the student selection test of the student selection and placement center.* Unpublished master's thesis, Middle East Technical University, Ankara, Turkey.

Levine, M. V. & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–290.

Li, C. H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 1538039469)

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

McDonald, R. P. (1981). The dimensionality of test and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525–542.

Muthén, L.K., & Muthén, B.O. (1998-2011). *Mplus 6*. Los Angeles, CA: Muthén and Muthén.

Önder, İ. (2007). Model veri uyumunun araştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *32*, 210–220.

Pelton, T. W. (2002). *The accuracy of unidimensional measurement models in the presence of deviations for the underlying assumptions.* Unpublished doctoral dissertation, Brigham Young University, Department of Instructional Psychology and Technology.

Progar, Š, Sočan, G., & Peč, M. (2008). An empirical comparison of item response theory and classical test theory. *Horizons of Psychology, 17*(3), 5–24.

R Core Team (2015). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Reise, S., & Waller, N. (2002). Item response theory for dichotomous assessment data. In Drasgow, F. and Schmitt, N. (Eds.), *Measuring and Analyzing Behavior in Organizations*. San Francisco: Jossey-Bass.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1–36.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63–84.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Teker, G. T., Kelecioglu, H., & Eroglu, M. G. (2013). An investigation of goodness of model data fit. *Procedia – Social and Behavioral Sciences, 106*, 394–400.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York: Teacher's College.

Van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*, 123–140.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33–45.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145.

UZUN ÖZET

Giriş

Eğitim ve psikoloji alanındaki test uygulamaları kişilerin bilgi, beceri ve tutum gibi örtük değerlerini en güvenilir biçimde ölçmeyi amaçlamaktadır. Bu tarz uygulamalarda standart testler en çok tercih edilen ölçme aletleri olmuştur. Standart testler, üniversitelere giriş sınavlarından ehliyet sınavlarına kadar birçok önemli alanda kullanılmaktadır. Bu tarz testlerde madde analizi ve kişilerin puanlarının hesaplanmasında iki farklı yaklaşım kullanılmaktadır. Klasik test teorisi olarak bilinen ve Türkiye'deki sınavlarda da yaygın olarak kullanılmakta olan yaklaşım kişilerin başarı, tutum, ya da diğer yeteneklerini temsil eden ham puanların hesaplanmasına dayalıdır. Bu yöntem her ne kadar kolay bir şekilde uygulanabilmesi nedeniyle tercih edilse de elde edilen madde istatistikleri ve ham puanların sınavda kullanılan sorulara ve soruları cevaplayan kişilerin oluşturduğu örnekleme bağlı olmasından ötürü geri plana düşmüş ve yerine bu sorunları içermeyen madde tepki kuramı ortaya çıkmıştır. Madde tepki kuramına göre bilgi ve beceri gibi örtük özelliklerin sınavda kullanılan sorular ve bu soruları cevaplayan kişilerin seviyelerinden bağımsız olarak ölçülmesi amaçlanmaktadır.

Madde tepki kuramının geçmiş yıllarda yurtdışında Test of English as a Foreign Language (TOEFL) ve The Certificate of English, Türkiye'de ise Öğrenci Seçme Sınavı (ÖSS) ve ÖZDEBİR gibi geniş ölçekli testlere uygulanabilirliği incelenmiştir. Bu çalışma, madde tepki kuramının geniş ölçekli standart testlere uygulanmasına yönelik farklı bir empirik örnek sunmayı ve bu kapsamda madde tepki kuramının uygulanmasında ortaya çıkabilecek sorunları değerlendirmeyi amaçlamaktadır. Bu amaçla Akademik Personel ve Lisansüstü Eğitimi Giriş (ALES) Sınavı tek boyutlu madde tepki kuramı modelleri doğrultusunda incelenmiştir. ALES sınavı geniş ölçekli bir test olup Türk üniversitelerine yapılan yüksek lisans ve doktora başvuruları ve üniversitelerdeki akademik personelin belirlenmesi gibi birçok önemli alanda kullanılmaktadır. Madde tepki kuramının ALES sınavına uyarlanabilirliği benzer şekildeki klasik test teoremine dayalı diğer geniş ölçekli standart testlere yönelik de fikir sağlayacaktır.

Yöntem

Bu çalışmanın örneklemi olarak ALES sınavının 2010 yılı güz döneminden elde edilen veriler kullanılmıştır. Sınava giren 142.178 kişi arasından rastgele seçim yoluyla seçilen 5000 kişinin sınavda 160 soruya verdikleri cevaplar tek boyutlu madde tepki kuramı modellerine göre incelenmiştir. ALES sınavı Sayısal 1, Sayısal 2 ve Sözel olmak üzere üç alt testten oluşmaktadır. Sayısal 1 ve Sayısal 2 testleri 40, Sözel testi ise 80 soru içermektedir. Sınavdaki tüm sorular çoktan seçmeli olup cevaplar doğru ya da yanlış olarak değerlendirilmiştir. Sınavda kişilerin her soru için sadece bir seçenek işaretlemeleri gerekmektedir. Ayrıca kişilerin cevaplarını bilemedikleri soruları boş geçebilmelerine izin verilmiştir. Toplam sınav süresi 180 dakikadır.

ALES sınavı verilerinin madde tepki kuramı doğrultusunda incelenmesi üç aşamadan oluşmaktadır. Birinci aşamada sorulara verilen cevaplar ve sınavdaki ham puanlar tanımlayıcı istatistikler aracılığıyla incelenmiştir. Bu analizlerin amacı ALES'te kullanılmış soruların zorluk ve madde ayırıcılık indekslerini incelemek ve uygun olmayan soruları tespit etmektir. İkinci aşamada ise madde tepki kuramına dair model varsayımları değerlendirilmiştir. Bu temel varsayımlar sınavın tek boyutlu olması ve soruların yerel bağımsızlığıdır. Sayısal 1, Sayısal 2, ve Sözel testlerinin tek boyutluluğu her bir teste tek boyutlu doğrulayıcı faktör analizi uygulanarak incelenmiştir. Bu modelin her bir test için uygun model uyum indeksleri vermesi testlerin tek boyutluluğunu göstermektedir. Soruların yerel bağımsızlığı özelliği ise sınavda alt ve üst %20'lik dilimde bulunan kişilerin sorulara verdikleri cevaplar arasındaki korelasyona bakılarak incelenmiştir. Eğer yerel bağımsızlık varsayımı doğru ise bu iki gruptaki kişilerin cevapları arasında yüksek korelasyon bulunmaması gerekmektedir. Bu iki varsayım haricinde sınav süresinin yeterliliği, sorulara verilen cevaplardaki tahmin oranı ve bir parametreli model için madde ayırıcılık indekslerinin eşit oluşu gibi ikincil varsayımlar da incelenmiştir. Üçüncü aşamada ise bir parametreli, iki parametreli ve üç parametreli lojistik madde tepki kuramı modelleri Sayısal 1, Sayısal 2 ve Sözel testlere sırasıyla uygulanmış ve en uygun model tespit edilmeye çalışılmıştır. Madde ve testlerin uyumu belirtilen modellere uyumu hem istatistiksel hem de grafiksel yöntemlerle incelenmiştir.

Sonuç ve Tartışma

Çalışmanın sonuçlarına göre Sayısal 2 testinin diğer iki teste göre daha zor olduğu ve bu testin yüksek başarılı ve düşük başarılı öğrencileri daha iyi ayırt ettiği görülmüştür. Ayrıca genel olarak sınava giren erkek katılımcıların Sayısal 1 ve 2 testlerinde daha başarılı olduğu, bayan katılımcıların ise Sözel testinde daha başarılı oldukları görülmüştür. Sınavda kullanılan tüm soruların yeterince düzeyde madde ayırıcılık gücüne sahip oldukları belirlenmiştir.

Model varsayımlarının incelenmesinde tek boyutlu doğrulayıcı faktör modelinin Sayısal 1, Sayısal 2 ve Sözel testleri için yüksek model uyum indeksleri verdiği belirlenmiş ve bu sonuçlar doğrultusunda ALES alt testlerin tek boyutlu olduğu sonucuna varılmıştır. Soruların yerel bağımsızlığının incelenmesinde ise Sayısal 1 ve Sayısal 2 alt testlerinde yer alan sorular arasında yüksek korelasyonlara rastlanmamış ve yerel bağımsızlık varsayımının geçerli olduğu görülmüştür. Sözel testinde ise özelikle paragraf tipi sorularda aynı paragrafa dair cevaplanması gereken soruların bazıları arasında yüksek korelasyonlar görülmüştür. Sınav süresinin yeterliliği Sayısal 1, Sayısal 2 ve Sözel testlerinin son beş sorusundaki boş cevap oranlarına bakılarak incelenmiştir. Bu incelemeye gore Sözel testinin son sorularının soruların zorluk oranlarından bağımsız olarak yüksek oranda boş bırakıldığı, Sayısal 1 ve Sayısal 2 testlerinde ise boş bırakılan soruların genelde sınavdaki zor sorular olduğu tespit edilmiştir. Diğer iki varsayım incelendiğinde ise ALES sorularının eşit madde ayırıcılık indekslerine sahip olmadığı ve sınava katılanların soruların doğru cevaplarını tahmin etme konusunda daha çok soruları boş bırakma davranışına gittikleri görülmüştür.

Bir, iki, ve üç parametreli modellerin ALES alt testlerine uyumu incelendiğinde Sözel ve Sayısal 2 testlerine en uygun modelin üç parametreli model olduğu, fakat Sayısal 1 testinde iki parametreli modelin daha iyi uyum gösterdiği belirlenmiştir. Model parametreleri incelendiğinde ise bir ve iki parametreli modeldeki parametrelerin ve bu modellerden elde edilen sınav puanlarının daha stabil olduğu tespit edilmiştir. Üç parametreli modelde özellikle tahmin (guessing) parametresinin hesaplanması esnasında boş soru sayısının da yüksek olması nedeniyle beklenmedik değerler tespit edilmiştir.

Bu çalışmanın sonuçlarına göre madde tepki kuramının üstün istatistiki özelliklerine karşın sınavlara uygulanması aşamasında karşılaşılabilecek olası sorunlara dikkat çekilmiştir. ALES ve benzeri özellikteki geniş ölçekli testlere madde tepki kuramının uygulanabilmesi için sınavların bu doğrultuda önceden dikkatle tasarlanması gerektiği görülmüştür.