International Journal of Engineering & Technology, *8 (1.4) (2019) 437-447*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Application of Alternative IRT models to IRT Assumption Violation

Yoon, Jiyoung[1] · Lee, Yoonsun[2]*

[1]The Korea Chamber of Commerce & Industry
[2] Seoul Women's University
**Corresponding author E-mail: ylee@swu.ac.kr*

Abstract

The purpose of this study is to investigate the most appropriate alternative IRT parameter estimation models among bi-factor model, testlet based model, and second-order IRT model when the IRT assumptions are not met. A simulation study was conducted to compare the alternative IRT parameter estimation models when the assumptions are unsatisfied. First, the comparison of the IRT models using the simulation data set without the satisfaction of the unidimensionality assumption indicated that bi-factor IRT model appeared the best fitting model. Second, when the local independency assumption was violated, the testlet based model appeared the best fitting model. The results of this study indicated it is necessary to estimate alternative IRT models by going through the process of anticipating the possibility of IRT assumptions violation due to the test forms and domains of their contents. This study also suggested that such a process will provide the basis for applying the IRT more precisely in order to estimate the capability of item and person characteristics

Keywords: *IRT assumption, MIRT, bi-factor model, testlet based model, second-order IRT model.*

1. Introduction

Large-scale assessment has been diverse in test purposes, content categories, items, and test takers; and they are used to assess and compare students' academic achievement levels. Moreover, large-scale assessment results can be generalized to a wide range of students; from the targeted students whose academic achievement levels are assessed to those from school level, local community level and even to national level. Thus, assessment results can be used widely that their effects can be great: (1) specifically, the test results help individual learners figure out appropriate educational goals and programs to improve their learning; (2) the differences in academic achievement levels among schools and local communities help diagnose regional differences in education and (3) also the test results help those in concern enhance the understanding of the differences among countries with different educational systems.

Due to the positive effects of large-scale assessments, various tests have widely been implemented in both domestic and global settings. For example, IEA (International Association for the Evaluation of Educational Achievement)'s TIMSS (the Trend in International Mathematics and Science Study) and OECD's PISA (the Program for International Student Assessment), were globally established to compare participating countries' educational levels, to manage the quality of their educational systems, and to figure out their students' achievement levels.

To compare participating countries' educational achievement levels using the test results, test validity and reliability should be presumed, and the accuracy in examining test items and subjects characteristics are required. A variety of measurement theories have been introduced to examine test and subject characteristics and item response theory (IRT) has been utilized to more precisely understand test items and subject characteristics (de Ayala, 2009).

Two assumptions to apply IRT to large-scale assessments are required: the first assumption is unidimensionality that a test should include only one dimension; the second is local independence that the observed items should be independent of each other (de Ayala, 2009). If the unidimensionality assumption is violated, the accuracy of estimated item parameter drops. Also, it negatively affects the interpretation of item bundle scores included in a test and test scores, test biasness, and test equating (Bolt, 1999; De Champalain, 1996; Oshima & Miller, 1992; Tata, 2004). If the local independence assumption is violated, estimated parameters become uncertain; test information function and reliability are overestimated; and estimation error is calculated extremely low (Demars, 2006; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Yen, 1993). In other words, if the assumptions are not satisfied, test items and subject abilities cannot be accurately estimated because items and subjects' abilities are not explained in a single dimension. Despite the problem, some preceding studies applied IRT without checking whether or not the assumptions were met (Rijmen, 2010; Jang, & Roussos, 2007; Wainer, & Wang, 2000; Yen, 1993).

The relationship between the two assumptions has been discussed in the previous studies. Tao (2008) presented that if the assumption of unidimensionality is violated, there is a probability of local independence violation (Tao, 2008). The reverse is true; if the local independence is not met, unidimensionality could be violated as well. Because the unidimensionality is about the number of dimensions within a test, the local independence violation caused by inter-item relationships could lead to the violation of unidimensionality. However, the two assumptions do not always take place together. In particular, previous studies argued that the inter-item responses clustered similarly because of test tools cannot be regarded as a specific dimension within a test (de Ayala, 2009; Goldstein, 1980). Based on controversial results, the present study separated the two assumptions as mutually independent.

Test contents is another factor to cause the violation of the IRT assumptions. Previous studies explained that reading tests violated the IRT assumptions frequently because one set of test includes multiple item bundles distributed to different reading passages (Choi, 2010; Pomemerich, & Segall, 2008; Schedl, Gor-don, Carey, & Tang, 1996; Wilson, 2000). Therefore, this present study adapted the characteristics of the reading test and set the form of tests including item bundles such as reading test as a special case with possible local independence violation.

Parameter estimation models, when the assumptions are violated, produced different outcomes. Previous studies developed and applied parameter estimation models applicable to the situations where the violation of the IRT assumptions takes place (Janssen, Tuerlinckx, Meulder, & de Boeck, 2000; Jiao, Kamata, Wang, & Jin, 2012; Jiao, Wang, & He, 2013; Rijmen, 2010). Jiao, et al. (2012) simulated large-scale data designed to include four sets of reading passages in the testlet-based, multi-level, and mixed models (combining the testlet-based model and multi-level model) for parameter estimation. Rijmen (2010) utilized reading test data consisting of four sets of passages and compared the bi-parameter IRT model assuming unidimensionality, hierarchical IRT model and bi-factor IRT model. Also, Wiberg (2012) applied the single-dimension IRT model and multi-dimensional IRT model for the IRT analysis of SAT implemented in Sweden. The studies found that the unidimensionality model showed the advantage of a simple model but the multi-dimension IRT model showed a higher goodness of fit.

As explained above, the previous studies contended that the IRT assumptions could be easily violated and, thus, an alternative parameter estimation model should be explored and applied, accordingly. Although the previous studies introduced various alternative parameter estimation models including the single dimension (two parameter) model, MIRT, bi-factor IRT model, hierarchical second order IRT model and testlet-based model, most previous studies compared results from two or three selected alternative parameter estimation models.

Therefore, this study seeks to apply and compare all of the alternative models presented from the previous studies as multi-dimensional IRT model, bi-factor IRT model, testlet-based model and second-order IRT model including single dimension IRT model, and ultimately to identify which of them is most suitable when assumptions are violated. Moreover, this study aims at presenting the necessary conditions for alternative parameter estimation model selection in diverse assumption violation situations by providing the criteria for an appropriate number of items or subjects. Furthermore, this study provides information helping accurate estimation of large-scale achievement test takers' ability so that those in concern can utilize more reliable and valid test results. This present study contributes to implementing accurate parameter estimation meeting purposes of test utilization, to comparing educational purpose, curricula, current achievement levels of learning based on test results.

2. Alternative IRT models

2.1. Unidimensional Item Response Model (UIRT)

In order to estimate item characteristics including item difficulty and discrimination, IRT model was introduced (Lord, & Novick, 1968). The UIRT as same as IRT model is designed to measure one latent variable as shown in Figure 1(Yi, 2005).

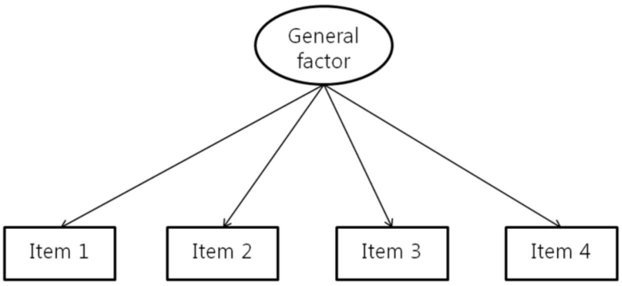

Fig. 1: Unidimensional IRT model

$$\rho(\vartheta) = \frac{e^{\alpha_j(\theta-\delta_j)}}{1+e^{\alpha_j(\theta-\delta_j)}} \quad (1)$$

In the equation (1) of UIRT model, the a represents item discrimination; θ, ability parameter; δj, item difficulty; and p(θ), the probability of answering the item correctly (de Ayala, 2009).

2.2. Multidimensional Item Response Model (MIRT)

Tests with more than two dimensions are defined to have the assumption of multi-dimensionality and MIRT model needs to be applied to those tests (Reckase, 2009). As in the equation below, the probability function is conducted the item discrimination parameter ($\boldsymbol{\alpha_{ik}}$), subject’s ability parameter ($\boldsymbol{\theta_{jk}}$), and difficulty position parameter ($\boldsymbol{\beta_i}$). The item difficulty parameter βi is defined with the difficulty parameter bi in the unidimensionality model as below (Reckase, 1985);

$$\rho\left(\chi_{ij} = 1|\theta_j, \alpha_i, \beta_i\right) = \frac{e(\sum_{k=1}^{k} \alpha_{ik}\theta_{jk}+\beta_i)}{1+e(\sum_{k=1}^{k} \alpha_{ik}\theta_{jk}+\beta_i)} \tag{2}$$

$$\beta_i = -\beta_i\sqrt{\alpha_{i1}^2 + \alpha_{i2}^2 + \cdots + \alpha_{ik}^2} \tag{3}$$

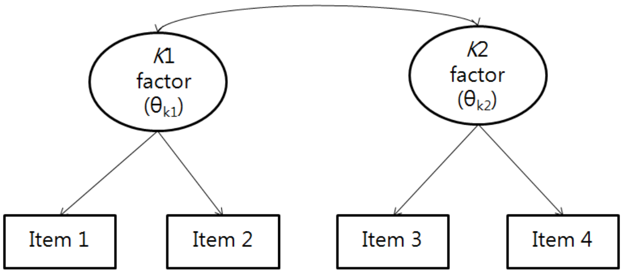

Fig. 2: Multidimensional IRT model

2.3. Bi-factor IRT model

With respect to the item characteristic analysis based on IRT, Gibbons & Hedeker’s study (1992) was the first to introduce the application of bi-factor model to IRT model under the item multi-dimensionality assumption. In the general dimension consisting of Thurstone’s single dimensions, the bi-factor model includes one or more different dimensions (Thurstone, 1947; Schmid & Leiman, 1957; Md Desa, 2012). The item response matrix reflecting it can be expressed as Figure 3.

In this bi-factor model, when item parameter is estimated, it is explained that items should follow general factors and include the additional second factor in a mutually exclusive way (Gibbons & Hedeker, 1992). In other words, the items included a specific content domain or cognitive domain in addition to the common factor as shown in Figure 4. When item parameter estimated in consideration of the specific content domain, the higher the accuracy of estimation is expected, compared with the case where item characteristics were estimated single factor assumption.

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & \alpha_{32} & 0 \\ \alpha_{41} & 0 & \alpha_{43} \\ \alpha_{51} & 0 & \alpha_{53} \\ \alpha_{61} & 0 & \alpha_{63} \end{bmatrix}$$

Fig. 3: The equation of full information bi-factor model

$$\rho(\mu|\theta) = \prod_{j=1}^{j} \rho(y_{j(k)}|\theta_g, \theta_k) \tag{4}$$

$$g\left(\pi_j\right) = \alpha_{jg}\theta_g + \alpha_{jk}\theta_k + \beta_j \tag{5}$$

In the equation (4) and (5), $\pi_j = p(y_{j(k)} = 1\left|\theta_g, \theta_k\right)$ means a latent variable explained probability equation (Rijmen, 2010). α_{jg} and α_{jk} represent inclination or the item j loading value of general latent variable(g) and particular latent variable(k). θ_g is a general latent variable applied to all of the items identically. θ_k refers to a particular latent variable applicable to a specific item group (Rijmen, 2010).

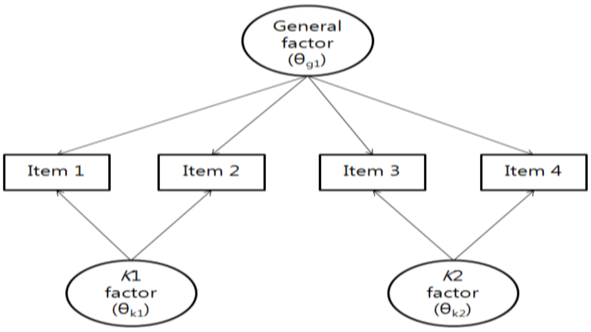

Fig. 4: Bi-factor IRT model

2.4. Testlet based Model

Testlet model is a special case of bi-factor IRT model (Rijmen, 2010; Wainer et al., 2007). This model refers to that a bundle of items with similar contents functions like a specific part of test (Wainer et al., 2007) as shown in Figure 5. It can be a form including multiple items in a reading passage. In the case of testlet-based test, if the reading passage is closely related with one's previous experience and that experience influences a response to items, the overall test efficiency and validity could be caused (Wang, Bradlow, & Wainer, 2003). Therefore, if the testlet-based test was estimated unidimensional IRT model, the item and person parameter has uncertain because of the testlet-based test is violated a local independence assumption.

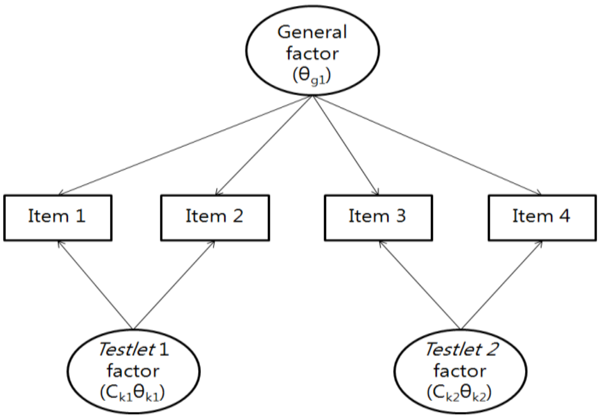
General factor (Θ_{g1})
Item 1
Item 2
Item 3
Item 4
Testlet 1 factor ($C_{k1}\Theta_{k1}$)
Testlet 2 factor ($C_{k2}\Theta_{k2}$)

Fig. 5: Testlet-based model

The testlet-based model can be explained in the equation below. In the equation, the specific constant C_k is fixed unlike the bi-factor model, meaning that the items belonging to the same group share particular resource (Rijmen, 2010). In the equation below, since C_k is a constant, it is noticed that $\alpha_{jg}C_k$ is identical to α_{jk} in the bi-factor model. The testlet-based model makes the items belonging to the same testlet be subject to the same constraint to estimate the probability.

$$g(\pi_j) = \alpha_{jg}(\theta_g + C_k\theta_k) + \beta_j = \alpha_{jg}\theta_g + \alpha_{jg}C_k\theta_k + \beta_j \tag{6}$$

2.5. Second-order IRT Model

Second-order IRT model can be explained as a model applying the IRT model to hierarchical factor model in Figure 6. The second factor model is similar to the bi-factor model in that it includes a general latent variable (g) and a particular latent variable (k). But it is different as its items are not directly affected by this general latent variable (Rijmen, 2010). This model that one item is directly affected by a particular latent variable and the particular latent variable is included in general dimension. That is, a particular latent variable is consequentially affected by general latent variable.

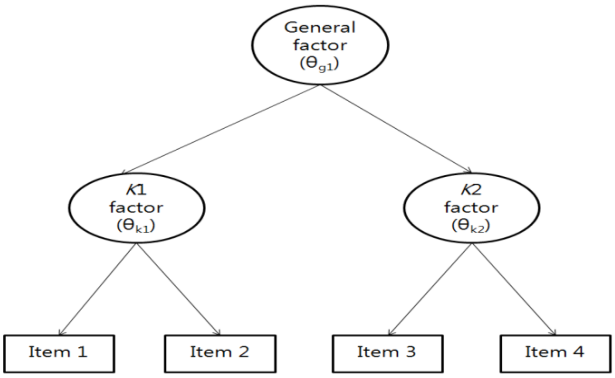
General factor (Θ_{g1})
*K*1 factor (Θ_{k1})
*K*2 factor (Θ_{k2})
Item 1
Item 2
Item 3
Item 4

Fig. 6: Second-order IRT model

The second-order IRT model appears as shown in the equation below; the particular latent variable, θ_k can be explained with particular latent variable (α_{kg}) and general latent variable (θ_g) refers to an error not explained by the general latent variable. Every particular latent variable is dependent on general latent variable and the error is statistically independent from the general latent variable (θ_g) (Rijmen, 2010).

$$g(\pi_j) = \alpha_{jk}\theta_k + \beta_j \tag{7}$$

$$\theta_k = \alpha_{kg}\theta_g + \xi_k \tag{8}$$

$$g(\pi_j) = \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\xi_k + \beta_j \tag{9}$$

By reflecting the particular dimension measures, the equation above is re-defined as follows; in the equation below, the general dimension is calculated by multiplying a particular constant (a_{jg}). Compared with the bi-factor IRT model, the second-order IRT model has a stricter restriction of bi-factor IRT model (Rijmen, 2010).

$$\begin{aligned} g(\pi_j) &= \alpha_{jk}\alpha_{kg}\theta_g + \alpha_{jk}\alpha_{jk}/\alpha_{kg}\xi_k + \beta_j \\ &= \alpha_{jk}\alpha_{kg}(\theta_g + \xi_k/\alpha_{kg}) + \beta_j \\ &= \alpha_{ia}^{\#}(\theta_a + C_k^{\#}\xi_k) + \beta_i \end{aligned} \tag{10}$$

3. Method

3.1 Simulation data

This study is used the theoretically optimal simulation data for assumption violation to check about a parameter estimation model. First, if unidimensionality is broken, multi-dimension model should be applied (Adams, Wilson, & Wang, 1997; Mair, 2007). Multi dimensionality refers to that one item includes one or more factors. To conduct multi dimensionality data, each subject has person parameter vector consisting of D number of latent variables. This vector is expressed in multi variate normal distribution consisting of the sum of vector matrix with mean=0 in D x D dimension.

Second, if local independence is unsatisfied, correlation between pairing items is used (Jannarone, 1986; Mair, et al., 2007). When inter-item correlation is 0, it means the items are independent; and if inter-item correlation is 1, it means the local independence is assumed highly strongly violated.

The simulation data consisted of 30 items and 1000 persons. Under the IRT assumption, data were organized according to unidimensionality violation and local independence violation. For dimensionality, data were organized in the condition including two dimensions. The correlation between each dimension was set at 0.1, if it is highly independent; 0.4, if intermediate; and 0.7, if highly dependent (Mair, et al., 2007). The item bundle was organized to include 30%, 70%, and 100% of the total number of items.

R package was used to generate the simulation data. eRm is an R program package made by Mair, et al. (2007). The module creating simulation data within the package bases on normal distribution N (0,1) to form the simulation data of θ and β. Simulation data are formed by first, randomly establishing the p matrix of probability model in n*k dimension and estimate p_{iv} probability. Next, the p_{iv} matrix is converted into X matrix coded with 0 or 1. Under the normal distribution assumption, of the randomly extracted p_{iv} is smaller than another extracted p'_{iv}, the value is converted into 0; and equal to or larger than p'_{iv}, 1.

3.2 Procedure of Study

Parameter estimation model comparison in the procedures as follows; first, simulation data were created under the conditions of unidimensionality violation and local independence violation. Second, parameter estimation model was applied to simulation data according to each condition, then the goodness of fit index was compared to find the most appropriate parameter estimation model. To apply to the data with unidimensionality violation, UIRT, bi-factor model, and second-order IRT model were. The simulation data with local independence violation, UIRT, bi-factor model, and testlet-based model were chosen for estimation. The alternative parameter estimation model estimated based on 2PL model.

For the parameter estimation model comparison, goodness of fit index of models estimated, item parameter estimation's standard error, and reliability of scale score were compared. The fit index is to check how fit each parameter estimation model is and reflects the difference between model-based estimated values and actual data values (Li, et al., 2012). To conduct it, differences among deviance (-2lnL), AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) were examined. Then, the standard deviation of item parameter estimation was compared based on the bias of item parameter. Lastly, the reliability of parameter estimations through each model was compared using marginal reliability.

3.3 Data Analysis

For parameter estimation, goodness of fit index was compared. The AIC and BIC indexes converted the standardized values of –2loglikelihood and can be expressed in the equation below; k means the number of parameters; and N, observed numbers (Md Desa, 2012).

$$\begin{aligned} AIC(k) &= -2I(\hat{\theta}) + 2k \\ BIC(k) &= -2I(\hat{\theta}) + k\log(N) \end{aligned} \tag{11}$$

Second, the standard deviation of item parameter estimation was tested. The IRT parameter estimation model processes estimation based on EM algorithm (Paek, & Cai, 2013). Item parameters estimated based on EM algorithm report standard deviation according to estimation. The fewer the standard deviations of item parameter estimation, the more accurate the parameter estimation is. Therefore, comparing the standard deviation of item parameter estimation was conducted to compare the accuracy of item parameter estimation.

Recently, various methods have been reported to examine the standard deviation of diverse item parameter estimation. In this study, SEM (Supplemented EM approach) method was employed to compare the standard deviation of item parameter estimation, which is capable of estimating standard deviations stably even when the number of observation is small though there are lots of items (Paek, & Cai, 2013). SEM calculates observed data information matrix based on missing data information (Orchard & Woodbury, 1972; Paek, & Cai, 2013).

$$I_0 = I_c - I_m = [I - I_m I_c^{-1}] I_c = [I - \triangle] I_c \tag{12}$$

The measured data information was calculated based on the difference between the information based on complete data and information of data including missing value. In the equation above, I_c represents the complete information matrix; and I_m, information matrix including missing value. Here, I represents square matrix; and $\triangle = I_m I_c^{-1}$, the part of missing value information. Based on it, the observed information matrix I_o was calculated and the square root of produced I_o^{-1} matrix's diagonal element was calculated parameter estimation standard deviation (Paek, & Cai, 2013).

Lastly, to compare the marginal reliability, it represents the accuracy of estimated person parameters and was utilized in this study to compare the reliability of scale score assuming multi dimensions such as testlet-based model, MIRT model, bi-factor IRT model and second-order IRT model. Scale score reliability can be inspected in the equation as follows; is the mean error variance of θ; σ2 (θ), total variance of θ; and reliability of $\bar{p}$ (Cai, 2013).

$$\overline{V}(\theta) = \sum_{s=0}^{s} V(\theta|s)p(s)$$

$$\bar{p} = 1 - \frac{\overline{V}(\theta)}{\sigma^2(\theta)} \tag{13}$$

4. Results

4.1 Parameter estimation model comparison in dimensionality violation

4.1.1 Fit index comparison

Depending upon the degree of unidimensionality violation, the goodness of fit of four models (UIRT, MIRT, bi-factor IRT model, and second-order IRT model) was compared as shown in Table 1 and Figure 7. Using the two dimension data, the degree of correlation between the two dimensions was set at 0.1 for very low correlation level; 0.4 for normal level; and 0.7 for high level. The results showed when the two dimensions were assumed independent and dimension correlation was 0.1, difference between MIRT model and bi-factor IRT model was verified. The inter-model –2lnL value difference was found statistically insignificant (Δdf=29, NS); thus, MIRT was the most suitable. At the level of 0.4 where the two dimensions show intermediate degree of dimension correlation, the bi-factor model was more suitable than MIRT model (Δdf=29, p<.05). This was also true at the level of 0.7 where the two dimensions have high correlation dimension (Δdf=29, p<.05). The RMSEA value of MIRT model was 0.15; and bi-factor model, 0.18 ~ 0.19.

Table 1: Fit index using unidimensionality violation data

Factor corr	Model	df	deviance	AIC	BIC	RMSEA
0.1	UIRT	939	35808.53	35928.53	36222.99	0.15
	MIRT	938	34602.90	34724.90	35024.27	0.15
	bifactor	909	34562.24	34742.24	35183.94	0.19
	second order	937	34643.06	34767.06	35071.34	0.19
0.4	UIRT	939	35529.19	35649.19	35943.66	0.15
	MIRT	938	34999.25	35121.25	35420.62	0.15
	bifactor	909	34949.84	35129.84	35571.54	0.18
	second order	937	34999.73	35123.73	35428.01	0.18
0.7	UIRT	939	35279.59	35399.59	35694.06	0.15
	MIRT	938	35190.23	35312.23	35611.60	0.15
	bifactor	909	35154.01	35334.01	35775.71	0.18
	second order	937	35190.32	35314.32	35618.60	0.18

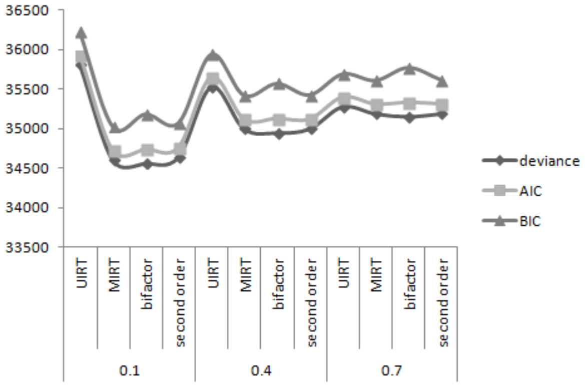

Fig. 7: Fit index

4.1.2 Bias comparison

After inspecting the optimal alternative parameter estimation model for the overall test structure of assumption violation data, this study sought to compare item parameter estimation accuracy at individual item level. To this end, parameter errors of items estimated through each model were compared. When estimating the item parameter of latent variable distribution through item response theory (IRT), the parameters estimated via EM algorithm should be reported together with their standard errors (Paek & Cai, 2013). Based on the comparison of standard errors of parameters estimated via each model, the alternative parameter estimation model examined based on the goodness of fit verification was assessed to see how accurately it estimated individual item difficulty and discrimination.

As a result of Kruskal Wallis difference verification, all of the parameters were found to have statistically significant differences as shown in Table 2. The standard errors of item parameter estimation were similar in their degrees on average. But the standard errors of parameter estimation by the bi-factor model were found higher than those of other models. Nevertheless, the standard errors of parameter estimation were the smallest in the hierarchical IRT model. It is because the bi-factor model includes more extreme values than other models.

Table 2: Standard error estimation using unidimensionality violation data

Factor corr.	Parameter	UIRT	MIRT	bi-factor	second order	Kruskal Wallis
.1	a	.10(.02)	-	.24(.07)	.07(.01)	79.14***
	a1	-	.12(.01)	.12(.01)	.07(.01)	26.84***
	a2	-	.12(.01)	.17(.03)	.06(.00)	45.08***
	d	.08(.02)	.09(.01)	.09(.02)	.09(.01)	11.57**
.4	a	.10(.01)	-	.14(.05)	.08(.01)	66.73***
	a1	-	.12(.01)	.13(.01)	.08(.01)	36.80***
	a2	-	.12(.02)	.19(.05)	.09(.01)	34.58***
	d	.27(.06)	.09(.02)	.09(.02)	.09(.01)	72.84***
.7	a	.10(.01)	-	.11(.03)	.10(.01)	12.18**
	a1	-	.11(.01)	.23(.15)	.10(.01)	35.49***
	a2	-	.11(.01)	.14(.02)	.10(.01)	34.65***
	d	.08(.01)	.09(.01)	.10(.06)	.09(.01)	12.75**

**p<.01, *** p<.001

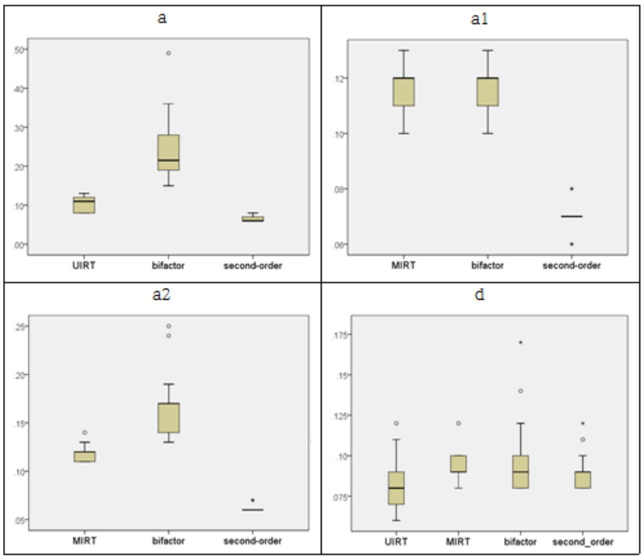

Fig. 8: Standard error estimation when factor correlation of 0.1

As in Figure 8, at the factor correlation level of .1, the range of parameter estimation by the bi-factor model ranged large. That is, the a, a1, and a2 parameter estimation error ranged larger in the bi-factor model. Generally, the standard errors of items estimated by the bi-factor model include more items seemingly extreme than those of other models. The d parameter also showed similar distribution but, in the case of the bi-factor model, the standard errors of item parameter estimation included extreme value items.

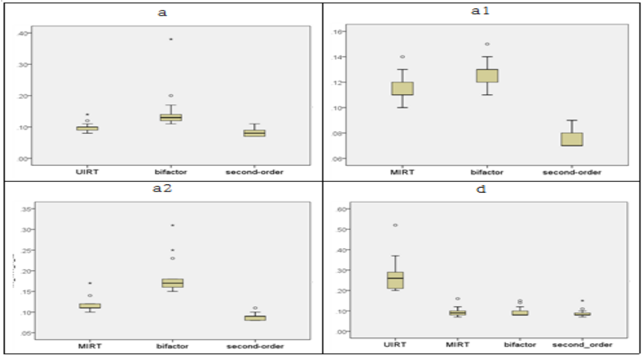

Fig. 9: Standard error estimation when factor correlation of 0.4

As in Figure 9, at the factor correlation level of .4, the range of parameter estimation by the bi-factor model gradually ranged narrower. However, as in the condition of inter-dimension correlation .1, a, a1, and a2 parameter estimation error ranged larger in the bi-factor model. The standard errors of items estimated by the bi-factor model included more items seemingly extreme than those of other models. The d parameter also showed similar distribution but the bi-factor model, the standard errors of item parameter estimation included extreme value items.

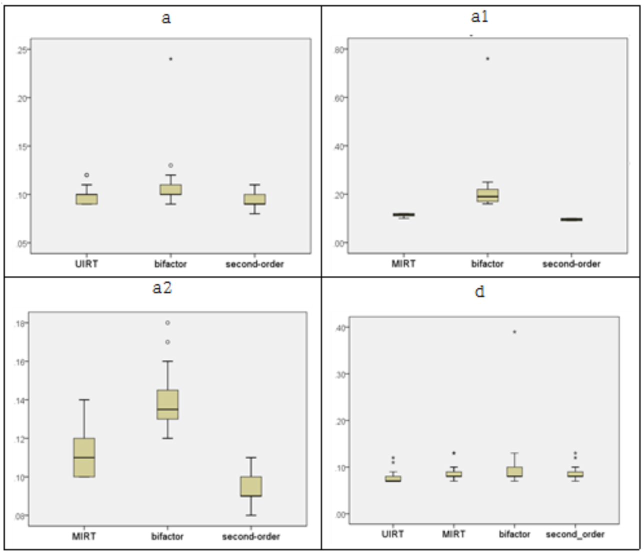

Fig. 10: Standard error estimation when factor correlation of 0.7

As in Figure 10, at the factor correlation level of .7, the parameter estimation of items estimated by the bi-factor model was found to range gradually narrower. However, as in the condition of the inter-dimension correlation .4, and .7, the bi-factor model's a, a1, and a2 parameter estimation error ranged larger. The standard errors of items estimated via the bi-factor model were found to include more items extreme than those of other models. The d parameter also shows similar distribution but the bi-factor model appeared that the standard errors of item parameter estimation included items with extreme values.

4.1.3 Marginal reliability comparison

Using the bi-factor model which is the optimal model for unidimensionality assumed violation data; the scale score reliability of other alternative parameter estimation models were compared. The results Table 3 showed that the reliability of measure point estimation of each model. The reliability of measure point estimation reported different patterns depending more upon the degree of correlation between the two dimensions, rather than the number of subjects. When the correlation between the two dimensions in this study was as low as 0.1, the reliability of bi-factor model's measure point was found very low; and when the correlation between the two dimensions was as high as 0.7, the reliability of the bi-factor model's measure point was high.

Table 3: Marginal reliability using unidimensionality violation data

Factor corr.	UIRT	MIRT	bifactor	second-order
0.1	0.54	0.40	0.06	0.30
0.4	0.78	0.44	0.45	0.44
0.7	0.80	0.38	0.72	0.71

4.2 Parameter estimation model comparison to local independence violation

4.2.1 Fit index comparison

Previous studies found that, if a test consisted of exams including item groups, it could violate local independence among the hypotheses to be met for IRT parameter estimation (Wainer, Bradlow, & Wang, 2007). For this reason, the degree of item group inclusion was adjusted to 30%, 70%, and 100% of the total items to identify the optimal parameter estimation model in diverse situations. As an IRT parameter estimation model for local independence violation data, the UIRT, the alternative parameter estimation model of bi-factor model, and testlet-based model were selected and their goodness of fit was invested.

For model goodness of fit, deviance, AIC, and BIC values were examined as shown in Table 4. The results showed that the optimal model was found to support the testlet-based model in the 30% of item groups (df=928, p<.001). Compared with the bi-factor IRT model, their inter-model –2lnL value difference was not statistically significant; thus, it was found that the testlet-based model is the optimal model (Δdf=9, NS). Likewise, in the 70% of item groups and in the 100% of item groups, difference in inter-model goodness of fit between the testlet-based model and bi-factor was not statistically significant (70%: Δdf=16, NS, 100%: Δdf=25, NS). The RMSEA range was found 0.27~0.32. Therefore, the testlet-based model is found to be an appropriate model.

Table 4: Fit index using local independence violation data

Item groups	model	df	deviance	AIC	BIC	RMSEA
30%	UIRT	933	27329.84	27449.84	27744.31	0.12
	bifactor	921	27008.65	27146.65	27485.28	0.21
	testlet	928	27014.03	27140.03	27449.22	0.21
70%	UIRT	933	27382.76	27502.76	27797.23	0.12
	bifactor	910	26421.65	26587.65	26994.99	0.20
	testlet	926	26443.89	26577.89	26906.71	0.19
100%	UIRT	933	27304.07	27424.07	27718.53	0.12
	bifactor	898	25999.71	26179.71	26621.41	0.18
	testlet	923	26024.38	26164.38	26507.92	0.18

4.2.2 Bias comparison

In this study, every alternative parameter estimation model is based on 2PL wherein the item parameters include difficulty and discrimination. Among the item parameters, they include the discrimination parameter, parameter and difficulty parameter, d parameter. Table 5 shows mean of errors in the condition of each item groups. As a result, statistically significant difference was found, violating local independence hypothesis. Therefore, non-parameter statistical method was employed to verify the difference in parameter estimation errors.

Table 5: Standard error estimation using local independence violation data

Item groups	Parameter	UIRT	bi-factor	testlet	Kruskal Wallis
30%	a	.12(.03)	.16(.08)	.15(.06)	5.00
	d	.11(.04)	.19(.26)	.15(.08)	4.39
70%	a	.13(.03)	.23(.11)	.20(.07)	31.07***
	d	.12(.04)	.33(.31)	.23(.14)	21.07***
100%	a	.13(.02)	.33(.29)	.53(1.46)	58.56***
	d	.11(.03)	.47(.42)	.67(1.99)	46.73***

*** p<.001

As in Figure 11, in the 100% of item groups, the parameter estimation of items estimated by the UIRT and testlet based model was found to range gradually narrower. Also, as in the condition of the 30%, 70%, the bi-factor model estimation of error ranged larger. The standard errors of items estimated via the bi-factor model were found to include more items extreme than those of other models. The d parameter also shows similar distribution but the bi-factor model appeared that the standard errors of item parameter estimation included items with extreme values.

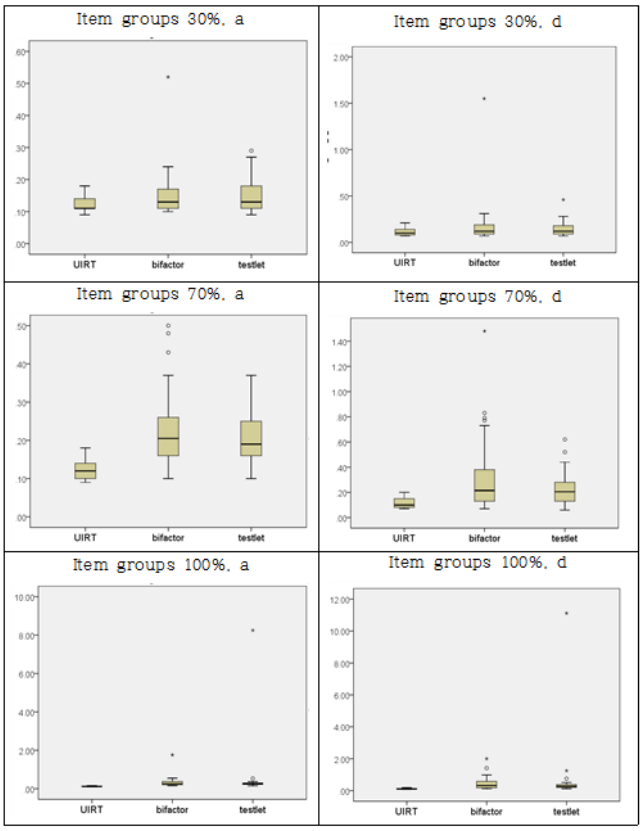

Fig. 11: Standard error estimation

4.2.3 Marginal reliability comparison

Based on the testlet based model optimal for the local independence violating data, the reliability of measure point estimated between other alternative parameter estimation models. Table 6 shows the reliability of measure point estimation of each model in the condition of 30% item groups. The marginal reliability shows when the test included 30% of item groups, the reliability of UIRT, bi-factor model, and testlet-based model's measure point ranged at a higher level of 0.77 ~ 0.80. The same result was found at the intermediate in the 70%, 100% of item groups as well.

Table 6: Marginal reliability using local independence violation data

Item groups	UIRT	bifactor	testlet
30%	0.80	0.77	0.77
70%	0.82	0.75	0.75
100%	0.81	0.73	0.74

5. Discussion

To identify the optimal alternative parameter estimation model in the IRT assumption violation, simulation data were created under the conditions with the number of items of 30 and the number of subjects of 1000. For unidimensionality assumption violation, tests with two dimensions were generated for multi-dimensional data. Inter-dimension correlation was set at 0.1, 0.4, and 0.7. For local independence violation, the degree of item group inclusion was set at 30%, 70%, and 100% of the total items.

UIRT, MIRT, bi-factor model, and second-order model were applied to the unidimensionality violation data to find the most suitable parameter estimation model. The results showed that at the inter-dimension correlation level of 0.1, the MIRT model and hierarchical IRT model were found suitable. However, at the inter-dimension correlation levels of 0.4 and 0.7, the bi-factor model was the most suitable. The results of the standard errors of item parameter estimation reported that the bi-factor model was found to have larger range of standard errors of parameter estimation model. This finding is identical to the results of Paek and Cai's study (2013) investigating how accurately an alternative parameter estimation model measured item parameters depending upon item length and number of subjects. They found that the errors of parameters estimated based on the bi-factor IRT model were larger than those via unidimensionality or two dimensional models. Moreover, they found that the larger the number of subjects and the longer the item length; the fewer the parameter estimation errors were. Likewise, in this study, the bi-factor IRT model was found to have more errors in item parameter estimation than the other two models; and the larger number of subjects, the smaller the item parameter estimation errors.

The reliability of test measure point was also examined. The findings revealed that when the inter-dimension correlation between the two dimensions included in the test was low at 0.1, the reliability of bi-factor IRT model's measure point was very low; whereas the reliability of measure point estimated via the MIRT model was high. On the contrary, when the inter-dimension correlation was high at 0.7, the reliability of bi-factor IRT model's measure point was found high.

The UIRT, MIRT, bi-factor model, and second-order model were applied to local independence violation data to identify the most suitable parameter estimation model.

For local independence violation, UIRT, MIRT, bi-factor model, and second-order model were applied to the data to identify the most suitable parameter estimation model. Regardless of item length, the most suitable model was the testlet-based model. When the testlet inclusion degree was 100%, rather than 30%, the standard errors of item parameter estimation increased generally in all of the models. This finding is consistent with Park's study (2010) applying testlet-based model and unidimensionality model to local independence violation test to examine their goodness of fit.

As the last criterion, the reliability of test measure point was investigated. The reliability of measure point in the three models showed no difference. Therefore, if local independence violation data are to be utilized to estimate measure point, the goodness of fit needs to be considered and find the most appropriate parameter estimation model to enhance the reliability of measure point estimation.

To summarize, in the IRT assumption violation situations identified based on simulation data, the minimum set of conditions to apply an alternative parameter estimation model are as follows; in the unidimensionality violation condition, the optimal alternative parameter estimation model is the bi-factor model. However, the bi-factor model, as found in the previous study, enlarges the standard errors of item parameter estimation and estimates more extreme values. Therefore, if the bi-factor model is employed for parameter estimation, this study suggests that the number of items should be considered when determining the number of subjects. In other words, in the case of large-scale assessment measuring students' academic achievement every year, since it tracks student development based on their measure point, it is necessary to verify the unidimensionality assumption; and, if the assumption is violated, consider a suitable number of subjects according to the number of items in each test and employ an alternative parameter estimation model for parameter estimation.

In the situation of local independence violation, the optimal model would be testlet-based model related to item composition. However, as a result of utilizing the testlet-based model for estimation, it appeared that the larger the item group inclusion in the overall test, the larger the standard errors of item parameter estimation. Tests such as PISA reading test, where the whole tests consist of item groups, will need to employ the testlet-based model in order to estimate item characteristics and subjects' ability.

Consequentially, if it is intended to consider test content dimensionality in using item response theory (IRT), bi-factor IRT model could be selected as an alternative parameter estimation model. When it is intended to consider the item group inclusion in a test, testlet-based model could be chosen as an alternative parameter estimation model. This study was implemented to provide the basic material for the determination of the optimal IRT model according to test composition. However, just as Rijmen's study (2010), it is necessary to consider factor composition in addition to the mathematical comparison. The hierarchical IRT model, of course, is a stricter model than the bi-factor model as it imposes more restrictions on coefficient. And a stricter model is feared to impact the goodness of fit depending upon the number of data. This present study also found a higher goodness of fit of other models in its mathematical comparison. However, in the practical aspect, when trying to select a model, the theoretic factor model of a test should be factored in along with the mathematical comparison of model goodness of fit in order to choose the optimal model. Nevertheless, as this present study proposed alternative parameter estimation models applicable to diverse IRT assumption violation cases, it is expected to serve as a guide for IRT model selection for accurate parameter estimation.

References

[1] Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinominal logit model. Applied Psychological Measurement, 21(1), 1-23.

[2] Bolt, D. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. Applied Measurement in Education, 12, 383–407.

[3] Cai, L. (2012). flexMIRT. Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric.

[4] Cai, L. (2013). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing (CRESST report 830). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standard, and Student Testing (CRESST).

[5] Choi, S. I. (2010). An application of full information item factor analysis to the reading comprehension of a TOEIC practice test. Journal of Educational Evaluation, 23(3), 709-734.

[6] de Ayala, R. J. (2009). The theory and practice of item response theory. New York: Guilford Press.

[7] De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. Journal of Educational Measurement, 33, 181–201.

[8] DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement, 43(2), 145-168.

[9] DeMars, C. E. (2010). Item response theory. New York: Oxford Press.

[10] Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. Psychometrika, 57, 423-436.

[11] Goldstein, H. (1980). Dimensionality, bias independence, and measurement. British Journal of Mathematical and Statistical Psychology, 33, 234-246.

[12] Jang, E. E. & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. Journal of Educational Measurement, 44(1), 1-21.

[13] Jannarone R. J. (1986). Conjunctive item response theory kernels. Psychometrika, 51(3), 357-373.

[14] Janssen, R., Tuerlinckx, F., Meulders, M., & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics, 25(3), 285-306.

[15] Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). Multilevel testlet model for dual local dependence. Journal of Educational Measurement, 49(1), 82-100.

[16] Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. Journal of educational measurement, 50(2), 186-203.

[17] Li, Y., Bolt, D. M., & Fu, J. (2006) A comparison of alternative models for testlets. Applied Psychological Measurement, 30, 3-21.

[18] Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test score. Reading MA: Addison-Wesley.

[19] Mair, P., & Hatzinger, R. (2007). Extended rasch modeling: the R package for the application of IRT models in R. Journal of statistical software, 20(9), 1-20.

[20] McDonald, R. P. (2000). A basis for multidimensional item response theory. Applied Psychological Measurement, 24, 99-114.

[21] Md Desa, Z. N. (2012). bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification. Unpublished doctoral dissertation, University of Kansas.

[22] Ochard, T., & Woodbury, M. A. (1972). A missing infromation principle: Theory and application. In L M. LeCam, J. Neyman, & E. L. Scott (Eds.), Proceedings of the sixth Berkerly symposium on mathematical statistics and probability (Vol. 1, pp. 697-715). Berkeley: University of California Press.

[23] Oshima, T. C, & Miller, M. (1992). Multidimensionality and item bias in item response theory. Applied Psychological Measurement, 16, 237–248.

[24] Paek, I., & Cai, L. (2013). A comparison of item parameter standard error estimation procedures for unidimensioanl and multidimensional item response theory modeling. Educational and Psychological Measurement, XX(X), 1-19.

[25] Park, C. (2010). A comparative study of IRT models for locally dependent reading test items by ESL leaners. Journal of Educational Evaluation, 23(2), 529-546.

[26] Pommerich, M., & Segall, D. O. (2008). Local dependence in an operational CAT: Diagnosis and implications. Journal of Educational Measurement, 45(3), 201-223.

[27] R Development Core Team. (2012). R: A language and environment for statistical computing[Computer software manual]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October, 10, 2014, from http://www.R-project.org

[28] Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9(4), 401-412.

[29] Reckase, M. D. (2009). Multidimensional item response theory. New York: Springer.

[30] Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the Testlet, and a Second-order multidimensional IRT model. Journal of Educational Measurement, 47(3), 361-372.

[31] Schmid, J. & Leiman, J. M. (1957). The development of hierarichival factor solutions. Psychometrika, 22(1), 53-61.

[32] Shedl, M., Gordon, A., Carey, P. A., & Tang, K. L. (1996). An analysis of the dimensionality of TOEFL reading comprehension items (TOEFL Research Report No. 53). Princeton, NJ: Educational Testing Service.

[33] Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. Journal of Educational Measurement, 28, 237-247.

[34] Tao, W. (2008). Using the score-based testlet method to handle local item dependence. Unpublished Doctoral Dissertation. University of Boston College. Department of Educational Research, Measurement, and Evaluation.

[35] Tate, R. (2004). Implications of multidimensionality for total score and subscore performance. Applied Measurement in Education, 17, 89–112.

[36] Thurstone, L. (1947). Multiple-factor analysis. Technical report, University of Chicago, Chicago, IL.

[37] Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Tests as an example. Applied Measurement in Education, 8, 157-186.

[38] Wainer, H., & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. Journal of Educational Measurement, 37, 203-220.

[39] Wainer, H., Bradlow, E. T., & Wang, X. (2007). Testlet response theory and its applications. New York, NY: Cambridge University Press.

[40] Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general bayesian model for testlets: theory and applications (ETS RR-02-02). Princeton, NJ: Educational Testing Service.

[41] Wiberg, M. (2012). Can a multidimensional test be evaluated with unidimensional item response theory? Educational Research and Evaluation: An International Journal on Theory and Practice. 18(4), 307-320.

[42] Wilson, K. (2000). An exploratory dimensionality assessment of the TOEIC test (ETS RR-00-14). Princeton, NJ: Educational Testing Service.

[43] Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. Journal of Educational Measurement, 30, 187-213.

[44] Yi, H. S. (2005). A method for estimating classification consistency of alternate forms under equating situations. Unpublished Doctoral Dissertation. University of Iowa. Department of Educational Measurement, and Statistics.

[45] Yoon, J. Y. (2017). Comparing alternative IRT parameter estimation models based on IRT assumption. International Journal of Internet of Things and Big Data. 2(2), 1-6. (Proceeding)