Journal of Education and Learning; Vol. 6, No. 4; 2017
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

Examination of Different Item Response Theory Models on Tests Composed of Testlets

Esin Yılmaz Koğar[1] & Hülya Kelecioğlu[2]

[1] Faculty of Education, Omer Halisdemir University, Niğde, Turkey

[2] Faculty of Education, Hacettepe University, Ankara, Turkey

Correspondence: Esin Yılmaz Koğar, Faculty of Education, Omer Halisdemir University, Niğde, 51200, Turkey. Tel: 90-388-225-4422. E-mail: esinyilmaz@ohu.edu.tr

Received: April 14, 2017 Accepted: May 16, 2017 Online Published: June 14, 2017

doi:10.5539/jel.v6n4p113 URL: http://doi.org/10.5539/jel.v6n4p113

Abstract

The purpose of this research is to first estimate the item and ability parameters and the standard error values related to those parameters obtained from Unidimensional Item Response Theory (UIRT), bifactor (BIF) and Testlet Response Theory models (TRT) in the tests including testlets, when the number of testlets, number of independent items, and sample size change, and then to compare the obtained results. Mathematic test in PISA 2012 was employed as the data collection tool, and 36 items were used to constitute six different data sets containing different numbers of testlets and independent items. Subsequently, from these constituted data sets, three different sample sizes of 250, 500 and 1000 persons were selected randomly. When the findings of the research were examined, it was determined that, generally the lowest mean error values were those obtained from UIRT, and TRT yielded a mean of error estimation lower than that of BIF. It was found that, under all conditions, models which take into consideration the local dependency have provided a better model-data compatibility than UIRT, generally there is no meaningful difference between BIF and TRT, and both models can be used for those data sets. It can be said that when there is a meaningful difference between those two models, generally BIF yields a better result. In addition, it has been determined that, in each sample size and data set, item and ability parameters and correlations of errors of the parameters are generally high.

Keywords: bi-factor model, item response theory, local dependency, testlet response theory

1. Introduction

Item Response Theory (IRT) is an influential theory presented as an alternative to the traditional true score model (Thissen & Orlando, 2001) and it is frequently used in the development and assessment of large scale tests due to its strong mathematical infrastructure. However, there are strong assumptions of IRT which are difficult to meet, and in order to determine the appropriate IRT model, first it should be assessed whether the data set at hand meets the IRT assumptions or not. One of these assumptions of IRT is the local independency. Local independency is the dependence of the possibility of providing correct answers to the items included in any measurement tool at a certain ability level on only the ability of the individual (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). However, this situation may be ruined in an item set consisting of items dependent on the same stimulant. Such items are defined as testlet by Wainer and Kiely (1987). Testlets defined to be longer than a single item but shorter than a whole test (Wainer, Bradlow, & Wang, 2007). There are studies in the related literature indicating that such items ruin the local independency assumption and are locally dependent (DeMars, 2006; Lee, 2004; Rijmen, 2010; Thissen, Steinberg, & Mooney, 1989; Zhang, 2010). Because the response of an individual to an item included in the testlet affects his/her responses to the other items in the same testlet (Wainer & Kiely, 1987). For this reason, local dependency emerges among the items depending on one stimulant.

There are many different reasons and types of Local Item Dependency—LID. Yen (1993) indicates ten different reasons for local item dependency: (a) external assistance or interference, (b) speededness, (c) fatigue, (d) practice, (e) item or response format, (f) passage dependence, (g) item changing, (h) explanation of previous answer, (i) scoring rubrics or raters, and (j) content, knowledge, and abilities.

Due to the content of large scale tests, studies mostly focused on passage dependency among these factors causing local item dependency; and it was identified in the conducted studies that local dependency was caused because the items were based on a common passage (Keller, Swaminathan, & Sireci, 2003; Lee, 2004; Sireci, Thissen, & Wainer, 1991; Thissen et al., 1989; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002). If there is more than one item depending on the same passage LID will occur. So that it is called passage dependence (Yen, 1993). That LID may be different based on the student's interests and background knowledge. The item set based on a common passage consisting of items related to each other depends on the individual's level of understanding of the passage. In this study, the main focus is on local item dependency that may arise from dependency on a passage. So, testlets based on common passage were examined in this study.

1.1 Approaches to Cope with Local Item Dependency in Testlet

Acting in accordance with the opinion that presenting stimulants to test takers takes much time in exams and, thus, asking one question in relation to a content will be a waste of time, test developers frequently construct testlets when developing tests which are based on a wide stimulant shared by an item set (Wainer et al., 2007). However, as the testlets violate one of the basic assumptions of IRT, different measuring models have been used in tests including a testlet in order to overcome the violation of that assumption. When the related literature is examined, the models used in the tests including testlets can be summarized as follows:

1.1.1 Dichotomous Item Response Models

For the cases where the effect of the testlets are ignored and items in the testlet are considered as independent items, Wainer and Lewis (1990) suggested using one of the two suitably categorized response models. In this model, the items included in a testlet are treated as any of the other items in the test (Arora, 2007; Wainer, Bradlow, & Du, 2000) and these are scored as independent units (Wainer et al., 2007). The data are analyzed using an appropriate model among 1PLM, 2PLM or 3PLM.

1.1.2 Polytomous Item Response Models

In another model, which is used to improve the local dependency among the items included in the testlets, the items in the testlet are accepted as a single item scored as multiple category items (Cook, Dodd, & Fitzpatrick, 1999; Sireci et al., 1991; Thissen et al., 1989; Wainer, 1995; Wainer et al., 2000; Yen, 1993). Five different items based on the same passage can be given as an example for this situation. In this approach, five items are not separately scored and not calibrated with the dichotomous IRT models. Instead, the set including five items is taken as a "single item" and examined with polytomous IRT models. In this manner, the score obtained as a result of polytomous scoring of the testlets is equal to the total true answers given (Arora, 2007; Wainer et al., 2000) and the testlet may get points ranging from zero to the total number of items related to the common stimulant (Boyd, 2003). The testlet score is indicated with the number of true answers (Wainer et al., 2000) as the items are independent among the testlets but dependent within a testlet (Wainer & Lewis, 1990), and by means of this scoring method, the dependency within the testlets can be eliminated (Dresher, 2002).

1.1.3 Bi-Factor Model (BIF)

The responses given to the testlets can be modelled with a multi-dimensional model and the two-factor model is appropriate for this content (DeMars, 2006). This model is a special condition of multi-dimensional IRT model and is developed by Gibbons and Hedeker (1992). In this model, each item is restricted in that it is weighted on two factors. In other words, in the bi-factor model, each item has a factor load not equal to zero in the primary factor and one of the secondary factors, and the factor load is zero for the other secondary factors.

1.1.4 Testlet Response Theory (TRT)

Another model used for the analysis of tests including testlets is the testlet effect model. This is the restricted state of the bi-factor model (Rijmen, 2009). The difference between this model and the bi-factor model is the inclusion of a model definition. Bi-factor model allows for separate differentiation parameters for the primary and secondary (testlet) dimensions, and those differentiation parameters can be independent of each other. For the analyses performed by using TRT, the scale of the secondary testlet dimension is not stable or the model definitions while the other restrictions of the bi-factor model (the location of all the dimensions and the scale of the general dimension are fixed) still exist for the testlet model (Rijmen, 2009).

In general, testlets have been treated as an independent item like the other items in the test and scored with dichotomous or polytomous standard IRT models, and various problems have been encountered. To explain them briefly, approaching testlets with standard models cause: (a) being partial in the determination of the difficulty of an item, (b) overestimation of the discrimination of items, (c) overestimation of the individuals' scores, (d)

overestimation of the test knowledge and the reliability, (e) errors in test equation results (Ackerman, 1987; Bradlow, Wainer, & Wang, 1999; Chen & Thissen, 1997; DeMars, 2006; DeMars, 2012; Lee, Dunbar, & Frisbie, 2001; Li, Bolt, & Fu, 2006; Sireci et al., 1991; Thissen et al., 1989; Tuerlinckx & De Boeck, 2001; Wainer, 1995; Wainer et al., 2000; Wang & Wilson, 2005; Yen, 1993). Especially the overestimation of individuals' scores increases the possibility of wrong classification, while categorizing the individuals based on their abilities may cause wrong decisions (Sireci et al., 1991; Yen, 1993). In their research, Tuerlinckx and De Boeck (2001) identified that the violation of the local dependency assumption caused partiality in the determination of the item and ability parameters. For this reason, correct modelling is of great significance for making the right decisions depending on the test results.

1.2 Purpose of the Research

The purpose of this research is to first identify (i) the item parameters obtained from the unidimensional item response theory model, bi-factor model and testlet response theory model used for estimating the parameters when the testlet number, independent item number and sample size change in the tests including testlets; and (ii) the standard error value belonging to these parameters; and (iii) the ability parameters and the standard error parameters belonging to these parameters and then to compare the obtained results.

This research is considered to be of significance for the related literature with respect to various aspects. The first reason stems from the recent increase of the usage of testlets in exams. Testlets are frequently used in standardized tests for efficient use of examination time by asking related questions (Wainer et al., 2000). It is accepted that items having a content dependent on each other are more appropriate for real life, and such items are regarded to measure high level abilities better (DeMars, 2006). For this reason, using testlets in the large scale tests is inevitable. Despite the advantages of using testlets in tests, violation of local independency, one of the most important assumptions of IRT, creates a negative situation. Therefore, the second reason making this research significant is the emphasis it lays upon the fact that standard IRT models ignore the local dependency among testlets and attract the attention to the communization of other models developed for testlets.

In many of the standardized education tests, there is an increase in the usage of testlets. For this reason, how to score and analyze tests consisting of testlets has been an important area of research in the last decade (Chien, 2008). Different opinions have been put forward in the scoring and evaluation of testlets. More complex models have been developed for tests consisting of testlets. In this way, local dependency among testlets have been taken into consideration, and despite not being compliant for such data sets, the usage of standard IRT models has tried to be avoided as testlets violate one of the most important assumptions of IRT. In a research study carried out by Min and He (2014), the model-data fit and parameter statistics obtained from the models dealing with testlet effect were shown to be better than the ones obtained from the unidimensional IRT (UIRT) model. In the analysis of the tests including testlets, more complex models such as the bi-factor model (BIF) are also used (DeMars, 2006; Min & He, 2014; Rijmen, 2010). In this research, testlet models are compared. Since this research draws the attention to the usage of more complex models according to UIRT, it is of significance to the literature from this aspect as well.

2. Method

2.1 Data Source

In this study, the data was obtained from the mathematic literacy cognitive test conducted in 2012 as one of the components of PISA (The Programme for International Student Assessment) which is a project of the OECD (Organisation for Economic Co-operation and Development). In the PISA applications, all of the students do not respond to all the items. For this reason, one of the 13 booklets was selected to work with the students having responded to the same items. In compliance with the purpose of the research, booklet number 10 was found to include many items and testlets, so it was selected as the data collection tool of this research. That booklet included a total of 36 items consisting of 14 independent items and 8 testlets. Of the 8 testlets in the booklet, 4 consisted of 2 items, 2 consisted of 3 items and 2 consisted of 4 items. In PISA 2012, there were 35545 students responding to the 10th booklet. The population of this research included three different samples in groups of 250, 500 and 1000 students randomly selected from the 28741 students remaining as a result of the elimination of data.

In the present research, 6 data sets were created using the testlets and independent items in booklet number 10. The number of independent items and testlets are provided in Table 1.

Table 1. The features of the data sets created in the present research

Items	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
Number of independent items	0	7	14	0	7	14
Number of testlets	4	4	4	8	8	8
Number of total items	11	18	25	22	29	36

When Table 1 is examined, the features of the different data sets comprised of 0/7/14 independent items and 4/8 testlets can be observed. The data sets including 4 testlets consisted of 2 testlets with 2 items, 1 testlet with 3 items and 1 testlet with 4 items.

32 items out of 36 in booklet 10 were scored as 0 or 1. The other 4 items were open-ended items and were scored partially (categories 0, 1, 2). Three of these four items were testlets (two included in the same and one included in a different testlet), the other was an independent item. The open-ended items in the scored dataset were encoded as zero for the wrong answers, as one for all the answers that received partial points, and as two for the ones that received full points.

2.2 Data Analysis

In the analysis of the data, the IRTPRO 2.1 (Item Response Theory for Patient-Reported Outcomes) program was used. This computer program was developed by Li Cai, David Thissen and Stephen du Toit in 2011.

2.2.1 Examination of the Local Independency and Uni-Dimensionality Assumption

In order to test whether the items in the testlet were local dependent or not, the local dependency X^2 (LD X^2) statistic, expressed in the related literature as a statistic easy to calculate and showing good performance, was used (Liu & Thissen, 2012). In cases where LD X^2 value is greater than 10 and when a major dependency is in question among the items, if there is a LD X^2 value of 5 to 10 between item pairs, medium level local dependency exists among the items and this situation may be an indicator of local dependency between the item pairs (Cai et al., 2011).

In data set 1, data set 2 and data set 3, there were LD X^2 values for 11 item pairs to be examined for 4 testlets, among which one testlet had 4 items (1 x $C_2^4 = 6$), 2 testlets had 2 items (2 x $C_2^2 = 2$) and one testlet had 3 items (1 x $C_2^3 = 3$). In dataset 4, dataset 5, dataset 6, there were 8 testlets in total, among which 4 testlets had 2 items, 2 testlets had 3 items and 2 testlets had 4 items. Thus, standardized LD X^2 values of the 22 item pairs in total were dealt with for all the sample sizes where 4 item pairs were used for 4 testlets with 2 items (4 x C_2^2), 6 item pairs for 2 testlets with 3 items (2 x C_2^3), 12 item pairs for 2 testlets with 4 items (2 x C_2^4). When the values of LD X^2 of the items included in the testlets in the datasets were examined, it was identified that local dependent item pairs existed in most of the datasets.

2.2.2 IRT Models Used

In the research, three models were used to make a prediction. 2PL/G-KPM was used for the unidimensional IRT model. And bi-factor and TRT model was used for multidimensional IRT model. The parameters of all IRT models were obtained using IRTPRO 2.1.

As IRTPRO 2.1 allows for simultaneous prediction with dichotomous and polytomous IRT models, 2PLM was used for the items scored in two categories and G-KPM was used for the items scored in multiple categories. The program draws different tables for the items analyzed using 2PLM and G-KPM. In all the datasets the items analyzed via the 2PL model separated the a and c parameters, and the standard errors belonging to these parameters were obtained. In the IRTPRO outputs, parameter a is the slope, and parameter c is the intercept, which is related to the threshold parameter "b" (b = -c/a). In addition, as slope-threshold form cannot establish correct generalizations in multidimensional models, slope-intercept form was used for the models $(a(\theta - b) = a\theta + c)$ (Cai et al., 2011). In MIRT, it is not certain whether or not parameter b is interpreted similarly as in UIRT (Min & He, 2014). In this study, instead of threshold parameter b, intercept parameter c was used as MIRT models were better generalized with it. For the items analyzed with G-KPM, a and c parameter values and standard errors belonging to these parameters were obtained.

In the IRTPRO 2.1 software, there are three prediction models for the parameters which are Bock-Aitkin Expectation-Maximization (BA-EM), Adaptive Quadrature and Metropolis-Hastings Robbins-Monro (MH-RM). In this study, the parameter predictions were made using the Bock-Aitkin marginal maximum possibility prediction. Because Cai et al. (2011) expressed that the Bock-Aitkin EM algorithm may result in more effective

predictions for the unidimensional and bi-factor IRT models. In order to calculate the expected ability scores of the individuals, a posteriori (EAP) method was used because the average is predicted impartially in this prediction method and little standard deviation values are obtained (Zimowski, Muraki, Mislevy, & Bock, 2003). In addition, this method can be used to make predictions even for individuals with zero and full scores (Embretson & Reise, 2000). In the whole calibration process, the overall and specific dimensions in bi-factor and the general dimension in UIRT and TRT are limited to have the standard normal distribution N (0,1). Thus, parameter estimates obtained from different IRT models are ensured to be at the same scale level (Li Y., Li S., & Wang, 2010).

In order to compare different models and identify which model fit the data, log-likelihood (-2Log-Likelihood, -2LL) values were estimated. With the help of the number of parameters to be predicted, the degrees of freedom at a 0.05 significance level were compared to the X^2 values in the table, and the models were examined to reveal whether or not they were different from each other at a significant level and which models complied best with the data. Finally, the correlation values belonging to the item and ability parameters obtained from different models and the standard error values belonging to these parameters were calculated using the Pearson correlation coefficient.

In order to compare different models and identify which model fit the data, log-likelihood (-2Log-Likelihood, -2LL) values were estimated. With the help of the number of parameters to be predicted, the degrees of freedom at a 0.05 significance level were compared to the X^2 values in the table, and the models were examined to reveal whether or not they were different from each other at a significant level and which models complied best with the data. Finally, the correlation values belonging to the item and ability parameters obtained from different models and the standard error values belonging to these parameters were calculated using the Pearson correlation coefficient.

3. Results

a_1: slope parameter belonging to the general dimension, a_2: discrimination parameter belonging to the secondary dimension, c: intercept parameter and SE: standard error; the item parameters and error prediction averages obtained from the items scored in dichotomous included in datasets 1, 2 and 3 are provided in Table 2.

Table 2. The item parameters and error prediction averages obtained from the items scored dichotomously

Data Set	Sample Size	Models	a_1	SE	a_2	SE	c	SE
Set 1 (4 testlets)	250	UIRT	1.61	0.32	-	-	-0.70	0.28
		BIF	1.79	0.56	0.88	0.87	-0.76	0.41
		TRT	1.85	0.45	1.85	0.45	-0.79	0.32
	500	UIRT	1.64	0.23	-	-	-0.47	0.18
		BIF	1.82	0.33	0.87	0.59	-0.56	0.30
		TRT	1.76	0.25	1.76	0.25	-0.52	0.17
	1000	UIRT	1.53	0.15	-	-	-0.58	0.12
		BIF	1.69	0.20	0.88	0.36	-0.67	0.22
		TRT	1.61	0.18	1.61	0.18	-0.61	0.13
Set 2 (4 testlets-7 independent items)	250	UIRT	1.60	0.33	-	-	-0.61	0.32
		BIF	1.66	0.40	0.61	1.13	-0.67	0.33
		TRT	1.65	0.34	1.92	0.39	-0.61	0.25
	500	UIRT	1.51	0.21	-	-	-0.55	0.20
		BIF	1.64	0.27	0.93	0.54	-0.61	0.24
		TRT	1.63	0.22	1.82	0.25	-0.59	0.17
	1000	UIRT	1.44	0.14	-	-	-0.61	0.12
		BIF	1.54	0.18	0.88	0.36	-0.66	0.17
		TRT	1.49	0.16	1.62	0.18	-0.62	0.12
Set 3 (4 testlets-14 independent	250	UIRT	1.54	0.25	-	-	-0.40	0.33
		BIF	1.60	0.38	0.69	3.10	-0.41	0.32
		TRT	1.61	0.28	1.89	0.33	-0.38	0.25

items)		UIRT	1.45	0.19	-	-	-0.36	0.19
	500	BIF	1.55	0.25	0.97	0.48	-0.38	0.22
		TRT	1.55	0.21	1.78	0.26	-0.37	0.16
		UIRT	1.41	0.13	-	-	-0.43	0.12
	1000	BIF	1.50	0.17	0.93	0.34	-0.46	0.16
		TRT	1.46	0.15	1.60	0.18	-0.42	0.11
Set 4 (8 testlets)		UIRT	1.59	0.29	-	-	-0.42	0.28
	250	BIF	1.76	0.62	0.71	0.97	-0.53	0.64
		TRT	1.83	0.42	1.83	0.42	-0.50	0.45
		UIRT	1.51	0.29	-	-	-0.35	0.33
	500	BIF	1.75	0.39	0.85	1.02	-0.51	0.42
		TRT	1.70	0.39	1.70	0.39	-0.46	0.31
		UIRT	1.44	0.15	-	-	-0.43	0.22
	1000	BIF	1.62	0.22	0.76	0.65	-0.54	0.30
		TRT	1.58	0.19	1.58	0.19	-0.49	0.28
Set 5 (8 testlets-7 independent items)		UIRT	1.51	0.32	-	-	-0.46	0.35
	250	BIF	1.64	0.54	0.69	1.03	-0,53	0.46
		TRT	1.62	0.41	1.72	0.47	-0.45	0.39
		UIRT	1.46	0.20	-	-	-0.41	0.20
	500	BIF	1.63	0.38	0.80	0.81	-0,51	0.24
		TRT	1.63	0.24	1.72	0.26	-0.49	0.22
		UIRT	1.40	0,16	-	-	-0.49	0.19
	1000	BIF	1.54	0.21	0.74	0.58	-0.55	0.23
		TRT	1.51	0.16	1.58	0.17	-0.54	0.14
Set 6 (8 testlets-14 independet items)		UIRT	1.51	0.33	-	-	-0.36	0.25
	250	BIF	1.61	0.47	0.67	0.99	-0.38	0.45
		TRT	1.55	0.40	1.64	0.48	-0.43	0.35
		UIRT	1.45	0.19	-	-	-0.32	0.19
	500	BIF	1.57	0.32	0.77	0.82	-0.37	0.33
		TRT	1.56	0.23	1.69	0.27	-0.35	0.22
		UIRT	1.40	0.13	-	-	-0.05	0.14
	1000	BIF	1.50	0.19	0.72	0.49	-0.42	0.20
		TRT	1.50	0.19	1.61	0.24	-0.41	0.20

When the item parameters and error prediction averages obtained from the items scored dichotomously as provided in Table 2 were generally interpreted, the following findings were reached. In different sample sizes, for the slope parameters obtained from different models and different datasets, the lowest average error value was obtained from UIRT in general; the average errors of prediction yielded by TRT were lower than that yielded by BIF; the average error of parameter c does not always show a general tendency. Especially in the sample size of 250, the results calculated for the different models changed. However, as the sample size increased, the error value averages obtained from the different models came closer; in the small sized samples the findings in BIF and TRT did not show a certain pattern. Moreover, as the number of independent items increased in the dataset, the average error values of the parameters tended to decrease. This situation may have stemmed from the effect of the increase in the number of items on the predictions. However, it can be said that item number and sample size mostly influence slope parameters, but are less influential on intercept parameters. By adding a testlet in a dataset, generally the error averages of the parameters increased, and the parameter showing minimum variance was parameter c. In the light of these findings, it can be expressed that a small sample size has an important effect on the error predictions in the models including a special dimension together

with the general dimension. Besides, it was also found that an increase in item number decreased the error average of parameter a of the general dimension.

a_1: slope parameter of the general dimension, a_2: discrimination parameter of the secondary dimension, ɣ: intercept parameter and SE: standard error, the average of the item parameters and findings related to the error averages for the items scored in polytomous in all the datasets are presented in Table 3.

Table 3. The item parameters and error prediction averages obtained from the items scored polytomously

Data Set	Sample Size	UIRT			BIF				TRT			
		a (SE)	ɣ1 (SE)	ɣ2 (SE)	a1 (SE)	a2 (SE)	ɣ1 (SE)	ɣ2 (SE)	a1 (SE)	a2 (SE)	ɣ1 (SE)	ɣ2 (SE)
Set 1	250	1.57 (0.35)	-1.37 (0.34)	-0.62 (0.27)	1.53 (0.14)	0.63 (0.20)	-1.35 (0.10)	-0.54 (0.12)	1.35 (0.14)	1.35 (0.14)	-0.99 (0.11)	-0.45 (0.12)
	500	1.20 (0.16)	-0.83 (0.14)	-0.46 (0.16)	1.20 (0.19)	0.47 (0.26)	-0.81 (0.13)	-0.41 (0.18)	1.25 (0.17)	1.25 (0.17)	-0.82 (0.13)	-0.39 (0.15)
	1000	1.31 (0.13)	-0.97 (0.11)	-0.49 (0.12)	1.27 (0.14)	0.39 (0.20)	-0.95 (0.10)	-0.47 (0.12)	1.34 (0.14)	1.34 (0.14)	-0.71 (0.11)	-0.42 (0.12)
Set 2	250	1.07 (0.19)	-0.74 (0.20)	-0.50 (0.21)	1.12 (0.66)	0.24 (1.25)	-0.74 (0.16)	-0.45 (0.53)	1.15 (0.34)	1.15 (0.34)	-0.73 (0.15)	-0.45 (0.40)
	500	1.14 (0.14)	-0.78 (0.13)	-0.48 (0.16)	1.18 (0.17)	0.45 (0.23)	-0.79 (0.12)	-0.43 (0.17)	1.17 (0.16)	1.17 (0.16)	-0.79 (0.10)	-0.44 (0.16)
	1000	1.23 (0.11)	-0.94 (0.10)	-0.52 (0.11)	1.29 (0.14)	0.48 (0.20)	-0.96 (0.11)	-0.46 (0.13)	1.29 (0.12)	1.29 (0.12)	-0.95 (0.09)	-0.45 (0.12)
Set 3	250	1.32 (0.47)	-1.38 (0.77)	-1.59 (0.62)	1.24 (0.47)	0.47 (0.82)	-1.20 (0.23)	-0.47 (0.42)	1.28 (0.69)	1.28 (0.90)	-1.21 (0.24)	-0.43 (0.56)
	500	1.13 (0.15)	-1.18 (0.17)	-0.68 (0.17)	1.19 (0.17)	0.49 (0.22)	-1.18 (0.14)	-0.62 (0.20)	1.20 (0.16)	1.20 (0.15)	-1.18 (0.15)	-0.61 (0.18)
	1000	1.23 (0.11)	-1.34 (0.12)	-0.64 (0.12)	1.26 (0.12)	0.44 (0.17)	-1.35 (0.11)	-0.59 (0.13)	1.36 (0.15)	1.43 (0.17)	-1.35 (0.12)	-0.53 (0.15)
Set 4	250	0.84 (0.19)	-0.68 (0.18)	-0.57 (0.21)	1.43 (0.80)	0.40 (1.10)	-0.73 (0.24)	-0.40 (0.59)	1.44 (0.27)	1.44 (0.27)	-0.78 (0.19)	-0.39 (0.24)
	500	1.18 (0.14)	-0.64 (0.21)	-0.58 (0.15)	1.28 (0.18)	0.40 (0.24)	-0.65 (0.12)	-0.48 (0.20)	1.28 (0.20)	1.28 (0.20)	-0.65 (0.10)	-0.49 (0.18)
	1000	1.27 (0.10)	-0.79 (0.10)	-0.62 (0.11)	1.36 (0.13)	0.38 (0.17)	-0.82 (0.09)	-0.55 (0.12)	1.35 (0.11)	1.35 (0.11)	-0.80 (0.08)	-0.56 (0.12)
Set 5	250	0.82 (0.19)	-0.65 (0.24)	-0.58 (0.21)	1.34 (0.99)	0.44 (1.34)	-0.66 (0.19)	-0.45 (0.72)	1.29 (0.76)	1.29 (0.76)	-0.58 (0.16)	-0.33 (0.58)
	500	1.16 (0.13)	-0.63 (0.14)	-0.59 (0.15)	1.25 (0.18)	0.42 (0.24)	-0.63 (0.12)	-0.51 (0.17)	1.29 (0.16)	1.29 (0.16)	-0.62 (0.09)	-0.50 (0.16)
	1000	1.23 (0.11)	-0.78 (0.13)	-0.64 (0.11)	1.34 (0.13)	0.44 (0.18)	-0.80 (0.09)	-0.56 (0.12)	1.34 (0.14)	1.34 (0.14)	-0.81 (0.12)	-0.55 (0.14)
Set 6	250	1.18 (0.19)	-1.04 (0.20)	-0.59 (0.23)	1.35 (0.54)	0.54 (0.85)	-1.03 (0.24)	-0.47 (0.47)	1.28 (0.43)	1.25 (0.49)	-1.02 (0.18)	-0.53 (0.24)
	500	1.13 (0.13)	-0.96 (0.14)	-0.72 (0.17)	1.21 (0.17)	0.45 (0.23)	-0.95 (0.14)	-0.65 (0.20)	1.22 (0.15)	1.23 (0.15)	-0.95 (0.12)	-0.65 (0.18)
	1000	1.22 (0.10)	-1.12 (0.11)	-0.70 (0.12)	1.30 (0.13)	0.45 (0.17)	-1.13 (0.11)	-0.63 (0.13)	1.31 (0.13)	1.33 (0.13)	-1.13 (0.11)	-0.63 (0.13)

When the findings of the datasets scored polytomously provided in Table 3 are generally interpreted, it is seen that in all the sample sizes of parameter a, the error parameters obtained from UIRT are less than those obtained from BIF and TRT. The error values of parameters a and ɣ decrease in all the three theories with an increase in

sample size. The decrease in the BIF and TRT error values was higher than that in UIRT error values. In small sample sizes, the findings were in favor of UIRT. However, with an increase in sample size, the error values came closer. The most optimum and decisive findings were obtained from the sample size of 1000. It seems that for small sample sizes, UIRT is the appropriate method to use. There was no clear finding in relation to the effect of increase in the independent item number or the testlet number in a dataset on the average error effect of the item parameters. BIF and TRT results obtained from different sample sizes and different datasets were identified to be close to each other. Using different models were identified to have little effect especially on ɣ parameter predictions.

Table 4 presents the findings relevant to the ability parameters and the standard error values of ability parameters.

Table 4. The ability parameters and error prediction averages

Data Set	250			500			1000		
	UIRT	BIF	TRT	UIRT	BIF	TRT	UIRT	BIF	TRT
	θ (SE)	θ (SE)	θ (SE)	θ (SE)	θ (SE)	θ (SE)	θ (SE)	θ (SE)	θ (SE)
Set 1	0.00 (0.44)	0.00 (0.48)	0.00 (0.47)	0.00 (0.44)	0.00 (0.46)	0.00 (0.47)	0.00 (0.45)	-0.01 (0.48)	0.00 (0.47)
Set 2	0.00 (0.37)	0.00 (0.38)	0.00 (0.38)	0.00 (0.38)	0.00 (0.40)	0.00 (0.39)	0.00 (0.39)	0.00 (0.41)	0.00 (0.41)
Set 3	0.00 (0.33)	-0.01 (0.43)	0.00 (0.40)	0.00 (0.34)	-0.01 (0.44)	0.00 (0.39)	0.00 (0.34)	-0.01 (0.41)	0.00 (0.35)
Set 4	0.00 (0.34)	-0.01 (0.38)	-0.01 (0.38)	0.00 (0.35)	0.02 (0.41)	-0.01 (0.40)	0.00 (0.36)	0.00 (0.36)	0.00 (0.36)
Set 5	0.00 (0.31)	-0.02 (0.49)	-0.02 (0.34)	0.00 (0.32)	0.00 (0.41)	-0.01 (0.41)	0.00 (0.32)	0.00 (0.38)	0.00 (0.32)
Set 6	0.00 (0.28)	0.00 (0.52)	-0.03 (0.41)	0.00 (0.29)	0.00 (0.44)	-0.14 (0.45)	0.00 (0.30)	0.00 (0.41)	0.00 (0.35)

In the light of the findings provided in Table 4, it was identified that in the ability predictions, UIRT resulted in a lower value of average error compared to those yielded by BIF and TRT, and TRT resulted in a lower average error compared to that of BIF. In the three models, the error averages of the general ability decreased with the increase in the independent item number in the dataset. Besides, the sample size was found to be non-influential when the number of testlets was four, and the error average for general ability decreased when the number of testlets was eight.

The next section will dwell on the findings in relation to the fit indices obtained from different models. Under all conditions, the models whose local dependency was taken into consideration showed a better model data compliance than UIRT. In general, there was not a significant difference between BIF and TRT; both models could be used for these datasets. When there is a significant difference between these two models, BIF generally provides better results. Even in small size samples, when there was local dependency among the items, BIF and TRT showed a better harmony which was a favorable finding. No generalization could be made on the effect of the number of independent items and the number of testlets on the harmony between model and data.

Table 5 presents the findings the correlations values for item parameters about different models. In each sample size, the correlation coefficients between BIF-TRT was found to be high. This situation shows that a similar prediction was made for the error values of the item parameters of these two models, and for this reason they can be exchanged in the predictions for the item parameters. For the slope and intercept parameters calculated from different datasets and different sample sizes, the slope parameter can be said to be more sensitive to the model and sample size, and the intercept parameter yields similar results under any condition. However, a conclusion could not be drawn about the relation of the number of items in the dataset or the number of independent items.

Table 5. Correlation values between item parameters and the standard error of these item parameters obtained from different models

		UIRT-BIF		UIRT-TRT		BIF-TRT	
Sample Size	Data Set	Slope (SE)	Intercept (SE)	Slope (SE)	Intercept (SE)	Slope (SE)	Intercept (SE)
250	Set 1	0.96 (0.47)	1.00 (0.84)	0.89 (0.65)	0.99 (0.89)	0.96 (0.88)	0.99 (0.97)
	Set 2	0.99 (0.59)	1.00 (0.87)	0.99 (0.79)	1.00 (0.74)	0.99 (0.86)	1.00 (0.92)
	Set 3	0.96 (0.54)	0.97 (0.97)	0.96 (0.53)	0.97 (0.94)	1.00 (0.97)	1.00 (0.97)
	Set 4	0.93 (0.52)	1.00 (0.85)	0.92 (0.59)	1.00 (0.87)	0.88 (0.87)	1.00 (0.96)
	Set 5	0.93 (0.55)	0.99 (0.87)	0.94 (0.53)	1.00 (0.80)	0.99 (0.86)	1.00 (0.92)
	Set 6	0.98 (0.62)	1.00 (0.78)	0.74 (0.95)	0.99 (0.66)	0.70 (0.92)	1.00 (0.97)
500	Set 1	0.95 (0.83)	1.00 (0.96)	0.98 (0.96)	1.00 (0.98)	0.98 (0.80)	1.00 (0.97)
	Set 2	0.97 (0.87)	0.99 (0.96)	0.98 (0.83)	1.00 (0.96)	0.99 (0.88)	1.00 (0.96)
	Set 3	0.98 (0.85)	1.00 (0.96)	0.97 (0.77)	1.00 (0.88)	0.99 (0.92)	1.00 (0.86)
	Set 4	0.93 (0.74)	0.98 (0.85)	0.95 (0.78)	0.99 (0.82)	0.93 (0.95)	1.00 (0.91)
	Set 5	0.96 (0.64)	0.99 (0.81)	0.98 (0.93)	0.99 (0.74)	0.99 (0.72)	1.00 (0.89)
	Set 6	0.97 (0.74)	0.99 (0.77)	0.98 (0.87)	0.99 (0.87)	0.99 (0.80)	1.00 (0.80)
1000	Set 1	0.93 (0.91)	1.00 (0.95)	0.99 (0.92)	1.00 (0.98)	0.96 (0.93)	1.00 (0.96)
	Set 2	0.96 (0.70)	1.00 (0.97)	0.99 (0.57)	1.00 (0.92)	0.97 (0.92)	1.00 (0.92)
	Set 3	0.97 (0.81)	0.99 (0.94)	0.98 (0.61)	1.00 (0.98)	0.97 (0.90)	1.00 (0.93)
	Set 4	0.96 (0.86)	0.99 (0.91)	0.98 (0.89)	1.00 (0.95)	0.98 (0.91)	1.00 (0.99)
	Set 5	0.97 (0.67)	0.99 (0.71)	0.98 (0.83)	1.00 (0.91)	0.99 (0.78)	1.00 (0.88)
	Set 6	0.92 (0.73)	0.97 (0.86)	0.97 (0.73)	0.99 (0.84)	1.00 (1.00)	1.00 (0.94)

Table 6 presents the findings correlation to the ability parameters and the error values of these parameters. According to Table 7, the correlation coefficients between the ability parameters obtained from all the models in all the sample sizes ranged between 0.990 and 1.000. This situation shows that the same findings would be reached in any model for any ability prediction. When the correlation coefficient between the error values of the ability parameters was examined, the correlation coefficient between UIRT-BIF was found to be between 0.971 and 0.999; and the correlation coefficient between UIRT-TRT was found to be within the range of 0.987 and 0.999. When the correlation coefficient between BIF-TRT was examined, it was found to be within the range of 0.973 and 0.999. As

the correlation coefficient between the ability parameters and the errors of these parameters was high, similar results were obtained for the prediction of ability parameter of all the models. For this reason, it can be concluded that the predictions for the ability parameter were not much influenced by the employed model, the sample size, the number of independent items and the number of testlets.

Table 6. Correlation values between the ability parameters and the standard error of these ability parameters obtained from different models

Sample Size	Data Set	UIRT-BIF		UIRT-TRT		BIF-TRT	
		Ability	SE	Ability	SE	Ability	SE
250	Set 1	0.994	0.984	0.995	0.997	0.990	0.990
	Set 2	0.999	0.981	0.999	0.999	1.000	0.984
	Set 3	0.999	0.997	0.999	0.996	1.000	0.998
	Set 4	0.999	0.991	0.999	0.992	1.000	0.999
	Set 5	0.999	0.996	0.999	0.994	1.000	0.996
	Set 6	0.999	0.998	0.999	0.987	1.000	0.987
500	Set 1	0.995	0.971	0.998	0.993	0.998	0.974
	Set 2	0.998	0.984	0.998	0.995	1.000	0.991
	Set 3	0.998	0.995	0.998	0.994	1.000	0.999
	Set 4	0.998	0.989	0.998	0.988	1.000	0.993
	Set 5	0.999	0.994	0.999	0.994	1.000	0.999
	Set 6	0.999	0.998	0.999	0.992	1.000	0.993
1000	Set 1	0.995	0.933	0.999	0.997	0.994	0.973
	Set 2	0.998	0.990	0.998	0.998	1.000	0.993
	Set 3	0.999	0.998	0.999	0.997	1.000	0.999
	Set 4	0.999	0.994	0.999	0.991	1.000	0.998
	Set 5	0.999	0.997	0.999	0.996	1.000	0.999
	Set 6	0.999	0.999	0.999	0.993	1.000	0.994

4. Discussion

As IRT is a strong mathematical model, today it is frequently used in education tests. In order to use this theory, there are assumptions that must be met by the data set and one of its most important assumptions is local independence. However, as a result of the use of testlets based on a common stimulant, the assumption of local independence is disrupted. Many studies in literature have shown that the disruption of local independence causes false predictions in the item and ability parameters obtained by standard IRT models (Ackerman, 1987; Chang & Wang, 2010; Eckes, 2014; Ip, 2000; Marais & Andrich, 2008; Monseur, Baye, Lafontaine, & Quittre, 2011; Reese, 1995; Wainer, 1995). In the present study, testlets were focused upon and data sets were analyzed with different IRT based models, with different sets of data from the items in the tenth booklet of PISA 2012, which measured mathematical literacy. To summarize the purpose of the study, we compared the item and ability parameters obtained as a result of analyzing six different data sets with UIRT, BIF and TRT models and the errors of these parameters.

Besides, the model-data fit indices were examined for each dataset; subsequently, the correlation values between the item and ability parameters obtained from these three models and the errors yielded by these parameters were calculated.

In the study, the average error value of the discrimination parameter obtained from UIRT was estimated to be lower under all conditions compared to those yielded by the other two models. In their study, Ackerman (1987), Yen (1993), and Chang and Wang (2010) also stated that if local independency assumptions were violated, the standard error of the discrimination parameter obtained with the IRT models would be less. So (2010) and Min and He (2014) also stated that when the tests including testlets were analyzed in unidimensional models, there was a great error in the slope parameter predictions. In the present study, it was observed that models taking item dependency into account calculate greater error predictions of discrimination parameters when compared to

unidimensional models. For this reason, based on the studies in literature it can be said that more accurate results of the slope parameters can be achieved by using BIF or TRT when local dependency is in question. In a study performed by Min and He (2014), UIRT, BIF and TRT models were compared in terms of item parameters, and no difference was found between the model selection and the prediction parameter estimation; in addition, it was observed that the intercept parameters were not affected by the local dependency among the testlets. In the present study, the intercept parameters obtained through UIRT were identified to be generally lower than those in other models. Similarly, Min and He (2014) found reported in their study that the intercept parameter estimations were influenced less by the model used when compared to the slope parameter predictions.

In the present study, it was seen that as the sample size increased, the error values of the item parameters got closer and generated similar results. Zhang (2010) used dichotomous and polytomous IRT models and testlet models in his study and concluded that sample size was influential on the analysis results for the three models; better results were obtained from bigger sample sizes. In addition, Ra (2011) conducted a study on testlets under simulation conditions of 1000 and 2000 individuals and stated that the RMSE values of the item parameters decreased in the sample size of 2000. In the present study, better results were also obtained for the greatest sample size of 1000 individuals for the predictions of item parameters.

It was observed that the error averages for general ability showed lower values for UIRT when compared to BIF and TRT for all the sample sizes and datasets. In their study, Chang and Wang (2010) compared the standard errors of the ability parameters obtained from the standard IRT model and TRT model over the real data set, and similar to the results of this research, they found that when the local independency assumption was not ensured among the items, the standard error of the ability parameter was predicted to be lower. Another study was conducted by Eckes (2014), who expressed in his study that in analysis results made with standard IRT models where the testlet effect was ignored, standard errors were predicted lower than actual. Based on the results of these research studies, having lower values of error averages of ability parameters obtained from UIRT when compared to the other two models can be due to the existence of local dependency among the items. For this reason, the findings regarding the error of the ability parameters in the study can be stated to be parallel with those reported in the related literature. However, Chang and Wang (2010) concluded that ability parameters were predicted higher when standard IRT models were used ignoring the local dependency, but in the present study, the findings regarding the ability parameters were close for all of the three models.

The findings of the present study, which revealed that the correlation between the item parameters and errors of items parameters obtained from different models was not low, are similar to the findings of the study performed by Bradlow, Wainer and Wang (1999), who used a real dataset. The researchers found that under independent model and testlet model conditions, the discrimination predictions were close to each other. In another study carried out by Min and He (2014), the correlation between intercept parameters obtained using 2PL-BIF and BIF-TRT models was estimated to be 0.98 for both. Again in the same study, the correlation values of the slope parameters among these models were found to be lower. In this study, the intercept parameter correlations were high for all models; and the correlation values for slope parameters were lower compared to the used model and the datasets.

The findings of the present study where the correlations of the predictions obtained from different models for general ability were high are consistent with findings reported by studies in the related literature. In his study conducted with the PISA results, DeMars (2006) made ability predictions using a two-factor model, testlet model, independent item model and multiple categorized item model, and estimated the correlations between the ability predictions obtained from different models to be close to 1.00. Eckes (2014), Min and He (2014) and Bradlow et al. (1999) also stated that the ability parameters were similar in their studies.

It has been identified that the models considering the local dependency under all conditions provided a better model-data harmony than UIRT. In general, there is no significant difference between BIF and TRT, and both models can be used for these data sets. When there is a significant difference between these two models, BIF can be said to provide better results in general.

As a result of the present study, the use of various local dependency indices to control the local independency of the items can be recommended, and if locally dependent items exist, instead of using standard IRT models, using the models considering the local dependency of the items can be suggested. In addition, as the BIF and TRT predictions provide high correlation, one of these two models can be preferred by the researchers.

The present study was carried out using real data sets, and there were few item pairs with local dependency at high levels. When real data sets with higher local dependent item pairs are used, a study can be conducted to see how the predictions change for the standard IRT models and the models considering the local dependency. This study was carried out using real data sets and there was no interference in the existing situation. For this reason, under similar

simulation conditions, the study can be repeated creating the desired environment. In this study, 2PLM was used for the items scored in two categories for all the models. Further studies can be conducted performing the analyses with 3PLM.

Acknowledgments

This paper was produced from the first author's doctoral dissertation. The first author wishes to express her gratitude to The Scientific and Technological Research Council of Turkey (TUBITAK) for doctoral scholarship support.

References

Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. Retrieved from http://files.eric.ed.gov/fulltext/ED284902.pdf

Arora, A. (2007). *Creating a TIMSS 2003 problem-solving scale and examining the problem-solving achievement of United States eight-grade students in TIMSS 2003* (Unpublished doctoral dissertation). Boston Collage, United States.

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

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168. https://doi.org/10.1007/BF02294533

Cai, L., du Toit, S. H. C., & Thissen, D. (2011). *IRTPRO: User guide*. Lincolnwood, IL: Scientific Software International.

Chang, Y., & Wang, J. (2010). *Examining testlet effects on the PIRLS 2006 assessment*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden. Retrieved from http://www.iea-irc.org/fileadmin/IRC_2010_papers/PIRLS/Chang_Wang.pdf

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289. https://doi.org/10.2307/1165285

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous item response theory models in the context of testlet scoring. *Journal of Outcome Measurement*, *3*(1), 1-20.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145-168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *3*6, 104-121. https://doi.org/10.1177/0146621612437403

Dresher, A. R. (2002). *The exmination of local item dependency of NAEP assessments using the testlet model* (Unpublished Doctoral Dissertation). University of Pittsburgh.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response. *Language Testing*, *31*(1), 39-61. https://doi.org/10.1177/0265532213492969

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

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika*, *57*, 423-436. https://doi.org/10.1007/BF02295430

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer. https://doi.org/10.1007/978-94-017-1988-9

Ip, E. H. (2000). Adjusting for information inflation due to local dependency in moderately large item clusters. *Psychometrika*, *65*, 73-91. https://doi.org/10.1007/BF02294187

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Applied Measurement in Education*, *16*(3), 207-222. https://doi.org/10.1207/S15324818AME1603_3

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from tests composed of testlets. *Educational and Psychological Measurement*, *61*(6), 958-975. https://doi.org/10.1177/00131640121971590

Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q_3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74-100. https://doi.org/10.1191/0265532204lt260oa

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3-21. https://doi.org/10.1177/0146621605275414

Li, Y., Li, S., & Wang, L. (2010). *Application of a general polytomous testlet model to the reading section of a large-scale English language assessment* (ETS RR-10-21). Princeton, NJ: Educational Testing Service.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*(8), 688-670. https://doi.org/10.1177/0146621612458174

Marais, I. D., & Andrich, D. (2008). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*, 105-124. Retrieved from http://scseec.edu.au/site/DefaultSite/filesystem/documents/Reports%20and%20publications/Archive%20Publications/Miscellaneous/ARC%20documents/ARC-Report09_Effects_of_Response_dependence.pdf

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*(4), 453-477. https://doi.org/10.1177/0265532214527277

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *4*, 131-155. Retrieved from http://orbi.ulg.ac.be//bitstream/2268/103137/1/IERI_Monograph_Volume04_Chapter_6.pdf

Ra, J. (2011). *Sensitivity of prior specification within testlet model* (Unpublished doctoral dissertation). The University of Georgia.

Reese, L. M. (1995). *The impact of local item dependence on some LSAT outcomes* (Report No. LSAC-R-95-02). Newtown, PA: Law School Admission Council, Inc.

Rijmen, F. (2009). *Three multidimensional models for testlet based tests: Formal relations and an empirical comparison* (ETS Research Rep. No. RR-09-37). Princeton, NJ: ETS.

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. https://doi.org/10.1111/j.1745-3984.2010.00118.x

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

So, Y. (2010). *Dimensionality of responses to a reading comprehension assessment and its implications to scoring test takers on their reading proficiency* (Unpublished doctoral dissertation). University of California.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, *6*(2), 181-195. https://doi.org/10.1037/1082-989X.6.2.181

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157-186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*(1), 1-14. https://doi.org/10.1111/j.1745-3984.1990.tb00730.x

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203-220. https://doi.org/10.1111/j.1745-3984.2000.tb01083.x

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden, & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Dordrecht, Netherlands: Kluwer. https://doi.org/10.1007/0-306-47531-6_13

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511618765

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*(2), 126-149. https://doi.org/10.1177/0146621604271053

Yen, W. M. (1993). Scaling performance assessments Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement*, *39*(4), 291-309. https://doi.org/10.1111/j.1745-3984.2002.tb01144.x

Zhang, O. (2010). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations* (Unpublished master dissertation). University of Florida.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [Computer Program]. Chicago: Scientific Software Corporation.