A Model Evaluation When Associations Exists Across Testlets under Small Testlet Size Situations

Ou Zhang
Research and Evaluation Methodology Program, College of Education
University of Florida
119G Norman Hall, Gainesville, FL 32611-7047
E-mail: zhango@ufl.edu

M. David Miller, PH.D.,
Research and Evaluation Methodology Program, College of Education
University of Florida
119A Norman Hall, Gainesville, FL 32611-7047

Mac Cannady
Educational Research, Measurement, and Evaluation, Lynch School of Education
Boston College
140 Commonwealth Ave, Chestnut Hill, MA 02467

Poster Session Presented at Annual Meeting of the National Council on Measurement in Education (NCME) annual meeting (2011) at New Orleans, LA

ABSTRACT

This study investigated the effectiveness of ability parameter recovery for two models to detect the influence of the association between testlets under the small testlet size situation. A simulation study was used to compare two Rasch type models, which were the Rasch tesetlet model and the Rasch subdimension model. The results revealed that the Rasch subdimension model performed better than the Rasch testlet model as the existence of between testlets association. The results also indicated that as the sample size increased, the discrepancies between model estimates and the real data set increased. The study concluded that using the Rasch subdimension model for testlet item analyses is efficient for small testlet size and non-adaptive typed tests when between testlets association exists. In sum, the Rasch subdimension model offered an advantage over the Rasch testlet model as it avoided standard error of measurement underestimation between testlets and better ability parameter estimations in the small testlet size situations.

Key Words: IRT, non-adaptive test, small testlet, model fit

INTRODUCTION

An item bundle or testlet, hence forward referred to as a testlet, is a scoring unit, a set of items following the same prompt, within a test that is smaller than the whole test (Wainer & Kiely, 1987). Items within testlets are locally dependent because they are associated with the same stimulus. Local item dependence is problematic because it introduces unintended dimensions into the test at the expense of the dimension of interest (Wainer & Thissen, 1996). The unintended dimensions present a threat to the reliability and validity of inferences from the test. This threat to test reliability and validity may result in a greater chance of misclassification when making decisions regarding examinee ability categorization (Sireci, Thissen, & Wainer, 1991; Yen, 1993). Thus, the challenge for test developers is not to eliminate the item dependencies within testlets, but rather to find a proper solution such that it does not impact the reliability and validity of inferences from the test.

In previous research, several models have been proposed as solutions. The testlet model (Wainer & Wang, 2000), was explicitly introduced to solve the problem of local item dependency within testlets. This testlet model includes a random effect parameter to model the local dependence among items within the same testlet. So, in addition to the overarching latent trait (i.e. the general ability), an additional latent trait (i.e. testlet effect) is also added to the testlet model; for each additional latent trait a random effect parameter is added to the model. One constraint of the testlet model is that all the latent traits in the model are required to be independent of one another. So, the testlet model assumes that not only the overarching latent trait is independent of testlet effects but also the test effects are independent of one another. This approach avoids the overestimation of the test reliability and test information so that the statistics of the Rasch testlet model consistently perform better than the standard Rasch model when testlets are present. However, in practice, the dependence between and among items can be even more complex. The National Board of Osteopathic of Medical Examiners (NBOME) offers computer-based COMLEX-USA exams online. This computer-based exam series is designed to assess the osteopathic medical knowledge and clinical skills considered essential for osteopathic generalist physicians to practice medicine without supervision. The COMLEX-USA level-2 exam consists of 350 items in 7 blocks including 141 independent items and 209 testlet items grouped in 95 testlets. The testlet sizes range from 2 to 4 items per testlet. There are five item types throughout the test: A -single item, D-single Item with graph, B-matching item, S-testlet item, and F-testlet item with graph. Among all five-item types, there are 3 different types of testlet items (i.e. B, S, and F). In the COMLEX-USA level-2 test the responses to the items within testlets are correlated because the items within each testlet share the same stimuli. Therefore, the assumption of local item independence is violated. However, there are two practical circumstances to note for data like NBOME COMLEX-USA exams. First, not only is there associations within each testlet, but also there are possible associations (denoted as *testlet correlation*) between two or more testlets. This is because some testlets may have similar item format (i.e. both belong to one of the testlet item types, like B, S, F) and they may share similar content subdomain. So, possible associations may exist between these testlet items even though they do not belong to the same testlet. Therefore, the associations of items may not only exist within testlet, but also may occur across testlets.

Second, in previous testlet research, in order to obtain illustrative results to support hypotheses, testlet sizes were usually 5 or more items (e.g., Adams, Wilson & Wang, 1997). Small and moderate testlet sizes (2-4 items) were rarely applied (e.g. Zhang, Shen, & Cannady, 2010). This is potentially problematic because in practice, like the NBOME COMLEX-USA exam, testlet sizes are often small.

Currently, the testlet model method is widely used for testlet analyses. Although the advantages and disadvantages of this approach has been discussed under small testlet size circumstance (Zhang, Shen, Cannady, 2010), further investigation regarding item associations across the testlets is still needed. First, in the Rasch testlet model, σ_θ^2 has to be set at unity for model identification (i.e. $\sigma_\theta^2 = 1.0$). One limit of the testlet model is that the model requires all the latent traits to be independent of one another.

$$\Sigma = \begin{bmatrix} \sigma_\theta^2 & 0 & \cdots & 0 \\ 0 & \sigma_{\gamma_1}^2 & \cdots & \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{\gamma_D}^2 \end{bmatrix} \tag{1}$$

This constraint is too restrictive to allow for possible item association between testlets. Therefore, further exploration of the between testlet association is impossible in the testlet model. Latent trait dimensionality misspecification may occur if associations exist between testlets (i.e. covariance).

The subdimension model (Brandt, 2007a, 2008) has been proposed to solve the between testlets item association issue. The subdimension model is based on the assumption that each person has an overarching ability in the measured dimension (denoted as main dimension), and testlet effects (denoted as subdimensions) are independent of main dimension but allows for possible subdimension associations by constraining the sum of the testlet effects (i.e. subdimension effects) to zero.

In accordance with previous testlet research, the purposes inherent to this study is exploring the consequences of variation in correlation between testlets on model fit, test reliability, and ability parameter recovery of the models under small testlet size circumstance. By looking for the trend of how changes in testlet factors affect different models' estimations and the test reliability corresponding to the models, a guide for model selection is expected to emerge. Furthermore, it will provide guidance for future improvements in the estimation of tests like the NBOME COMLEX-USA exam.

THEORETICAL FRAMEWORK

Rasch Testlet Model

The Rasch testlet model includes a random effect parameter, which models the local dependence among items within the same testlet (e.g. Wang & Wilson, 2000). It can be written as

$$P_{ji1} = \frac{\exp(\theta_j - b_i + \gamma_{d(i)j})}{1 + \exp(\theta_j - b_i + \gamma_{d(i)j})} \qquad (2)$$

where P_{ji1} is the probability that examinee j answers item i correctly;

$\theta_j \sim N(0,1)$ is the ability of examinee j;

$b_i \sim N(\mu_b, \sigma_b^2)$ is the difficulty of item i, and

$\gamma_{d(i)j} \sim N(0, \sigma^2_{\gamma_{d(i)}})$ is a random effect that represents the interaction of person j with testlet $d(i)$ (i.e., testlet d that contains item i).

With j=1,…, J and J the total number of examinees, Restriction 1: $\sigma(\theta_J, \gamma_{jd(i)}) = 0$ for all $d = 1, ..., D$ (3)

Restriction 2: $\sigma(\gamma_{jd(i)}, \gamma_{jd(i')}) = 0$ for all $d = 1, ..., D$ (4)

Restriction 3: $\sum_{j=1}^{J} \theta_j = 0$ (5)

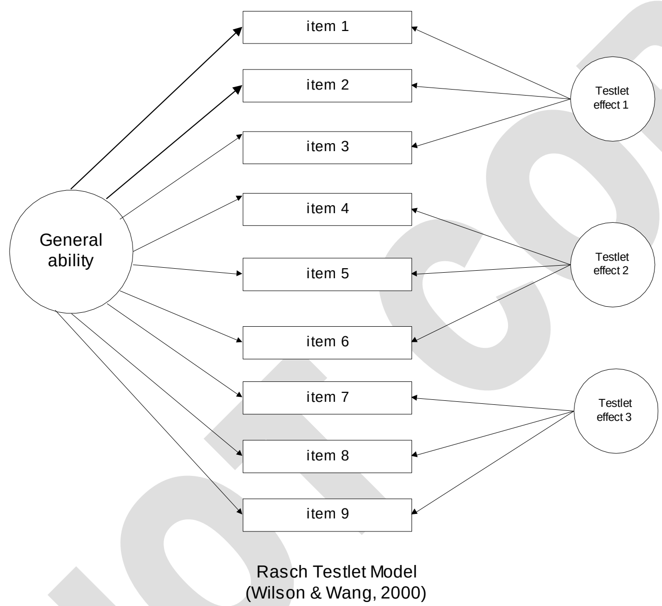

Figure 1. Rasch testlet model

<u>Rasch Subdimension Model</u>

Brandt (2007a, 2008) proposed the Rasch subdimension model, which is similar to the Rasch testlet model (Wang & Wilson, 2005) in that it allows for association between testlet effects. It can be written as follows:

$$P_{ji1} = \frac{\exp(\theta_j - b_i + \gamma_{d(i)j})}{1 + \exp(\theta_j - b_i + \gamma_{d(i)j})} \qquad (6)$$

where all the parameters in the model have the same definitions as the Rasch testlet model except Restriction 2.

Restriction 1: $\sigma(\theta_J, \gamma_{jd(i)}) = 0$ for all $d = 1,\ldots, D$ (7)

Restriction 2: $\sum_{d=1}^{D} \gamma_{jd(i)} = 0$ for all $j = 1,\ldots, J.$ (8)

Restriction 3: $\sum_{j=1}^{J} \theta_{jd} = 0$ (9)

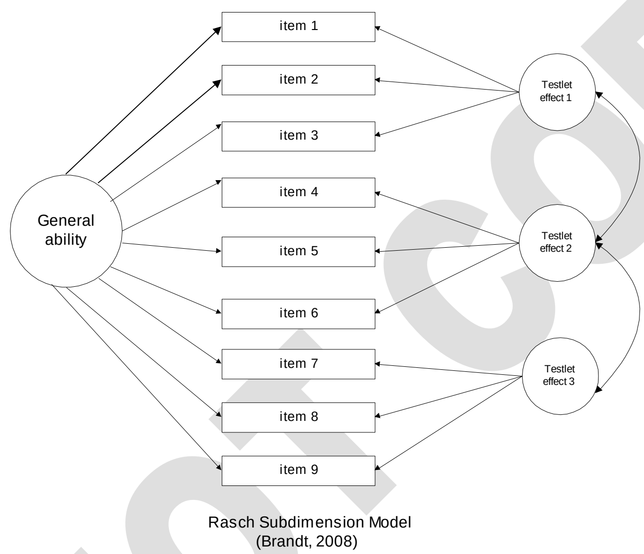

Figure 2. Rasch subdimension model

PURPOSE OF THE STUDY

In the past, very little research has focused on the variation of associations between testlets, desipte this phenomenon frequently occurring in realistic situations. Small testlet size issues have also been insufficiently explored in past research. The purposes of this study, therefore, are to explore the consequences of variation in between testlet item correlation on model fit, test reliability, and ability parameter recovery of the models under small testlet size circumstance. In order to provide guidance to model selection in testlet analyses under these two circumstances, comparisons between the Rasch testlet model and Rasch subdimension model are investigated. Furthermore, the result of the model assessment is displayed as guidance for future COMLEX-USA and other exam systems innovation.

METHOD

The Rasch testlet model and Rasch subdimension model are applied in this study. First, a series of simulation studies designed to investigate the extent to which the fluctuation of conditions influenced the different model fitting results, including the association between testlets and local dependence effects within testlets, were conducted. These simulations were conducted to evaluate model fit, test reliability, and parameter recovery of the two different IRT models. Next, the two models were fit to the COMLEX-USA exam dataset to investigate how they compare in an empirical case.

Model Used to Generate Data for the Simulations

The current study evaluated the effect of changes in the association between testlets and the local effect of testlets on the model fit, ability parameter recovery, and test reliability of different IRT models. In order to quantify the extent of these variations local effect, the application of Rasch subdimension model (Brandt, 2007a, 2008) was appropriate for the data simulation. The Rasch subdimension model included correlations between subdimensions and testlet effects ($\gamma_{d(i)j}$) within every testlet. Therefore, the Rasch subdimension model was used to generate data.

The following prior model constraints were used to simulate the responses.

With j=1,…, J and J the total number of examinees,

$$\sum_{d=1}^{D} \sigma_{\gamma_{jd(i)}\gamma_{jd(i')}} = 0 \text{ for all } j = 1,\ldots,J. \qquad (10)$$

$$\sigma(\theta_j, \gamma_{jd(i)}) = 0 \text{ for all } d = 1, \ldots, D \tag{11}$$

<u>Main Dimension and Subdimension Covariance Matrix Definition</u>

The main dimension and subdimension correlation matrix was defined according to the definition of the Rasch subdimension model (Brandt, 2007a, 2008). In this covariance matrix, the main dimension (i.e. σ_θ^2) was defined to be uncorrelated with any subdimensions. σ_θ^2 had to be set at unity for model identification (i.e. $\sigma_\theta^2 = 1.0$). Therefore, no non-zero off-diagonal component existed in the first column and row (i.e. covariance regarding main dimension). For the subdimensions, the covariance of the subdimensions might differ from zero according to realistic circumstances. However, based on the definition of the Rasch subdimension model, the last subdimension was set to cancel out all the subdimension covariance in the model. The subdimension model covariance matrix is shown as below.

$$\Sigma = \begin{bmatrix} \sigma_\theta^2 & 0 & 0 & 0 \\ 0 & \sigma_{\gamma_1}^2 & \sigma_{\gamma_1}\sigma_{\gamma_2} & 0 \\ 0 & \sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma_{\gamma_2}^2 & 0 \\ 0 & -\sigma_{\gamma_1}\sigma_{\gamma_2} & -\sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma_{\gamma_3}^2 \end{bmatrix} \tag{12}$$

<u>Data Source and Population parameters</u>

The population item parameters and ability parameters were randomly drawn from normal distributions for each condition (i.e. $\theta_j \sim N(0,1)$ $b_i \sim N(0,1)$). The response data were generated using the statistical software R 2.12.2. 100 sample response data were generated for each condition.

Parameter Estimation

In the study, the parameters of the dataset in 2 models were analyzed using Marginal Maximum Likelihood (MML) methods with ConQuest Version 2.0. The estimations of the simulees' abilities were calculated by Expected a Posteriori Estimation (EAP; Bock & Mislevy, 1982).

Simulation Design

Our study was a four-factor completely crossed design: 3 (testlet correlation changes) $\times 4$ (levels of local dependence effect) $\times 3$ (ratio of testlet items and independent items)$\times 2$ (sample size).

i. The testlet sizes chosen were based on the sizes less often discussed in the applied literature. Thus, for the simplicity of the study, only one testlet size (testlet size: 5) was used.
ii. Three different testlet correlations between similar testlet formats (i.e. B, S, F types) were applied (i.e. 0.1, 0.2, 0.3).
iii. The ratio of the correlated/total testlet numbers is very important in research. However, for this simplicity of the study, only three correlated testlets were included in this study. So, only one pair of positively correlated testlets was included,however, one negative correlated testlet was used to cancel out the association between aforementioned two testlets.
iv. Four levels of local dependence effect were examined: $\sigma_\gamma^2 = (0.25, 0.5, 0.75, 1)$
v. Among all 60 items, the ratio of testlet items to independent items were $(1:3, 1:1, 3:1)$
vi. Two different sample sizes of examinees ($500, 1000$) were applied.

ANALYSIS

In this study, each simulated data set was generated using ConQuest 2.0. The outcomes of interests were the model fit index- the likelihood ratio test ($-2 \times \log(likelihood)$) for comparing the deviance between the Rasch subdimension model and the Rasch testlet model. Since these two models are not nested, Akaike information criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC) were also calculated. The accuracy of estimation for item parameters and ability parameters was quantified via bias and root mean square error (RMSE) across all replications.

The likelihood ratio test ($-2 \times \log(likelihood)$) is a measure of the difference between the null model and alternative model. The likelihood ratio test is distributed as a chi-square statistic with degrees of freedom $df_D = df_{null} - df_{alt}$. Models with fewer parameters (e.g., the null model) are hypothesized to have larger loglikelihood than models with more parameters (alternative model). The likelihood ratio test is mathematically defined as:

$$\chi^2(df_D) = -2[\ln L_{null} - \ln L_{alt}] \tag{17}$$

Akaike's information criterion (AIC), is a measure of the goodness of fit of an estimated statistical model. AIC is mathematically defined as:

$$AIC = -2\ln L + 2P \tag{18}$$

where P is the number of estimated parameters. The model with the smallest AIC is the one to be selected.

Bayesian Information Criterion (BIC), is also a measure of the goodness of fit of an estimated statistical model and tends to favor more parsimonious models than the AIC. BIC is mathematically defined as:

$$BIC = -2\ln L + 2P\ln(N) \tag{19}$$

where P is the number of estimated parameters, N is the sample size. The model with the smallest BIC is the one to be selected.

Bias is defined as average difference between true and estimated parameters across all people and items. An estimate of bias is calculated for each replication under each condition giving an average bias of each condition in the simulation. Bias is mathematically defined as:

$$bias_\theta = \frac{\sum_{j=1}^{n} \hat{\theta}_j - \theta_j}{n} \tag{19}$$

where the θ_j is the true value of a item or person parameter;

$\hat{\theta}_j$ is the estimated value of that parameter ;

n is the total instances of that type of parameter within a replication (i.e. sample size for ability θ).

RMSE is a measure of absolute accuracy in parameter estimation. RMSE is calculated for each parameter type in a replication and an average for each condition is determined. RMSE is the square root of the average squared difference between estimated and true parameters, and is mathematically defined as:

$$RMSE_{\lambda} = \sqrt{\frac{\sum_{j=1}^{n}(\hat{\theta}_j - \theta_j)^2}{n}} \tag{20}$$

where terms in the equation are defined as they are with bias.

Reliability

In this study, test reliability coefficients were computed for item responses scored dichotomously for both Rasch testlet model and Rasch Subdimension model. As we use MML estimation in ConQuest, the test reliability can be calculated as

$$Test\ \mathrm{Re}liability = \frac{Var\,(\theta_T)}{Var\,(\theta_{EAP})} = \frac{S^2(\hat{\theta}) - (\overline{s.e^2_{\hat{\theta}}})}{S^2(\hat{\theta})} \tag{21}$$

RESULTS

Model Deviance, AIC and BIC

The magnitude of the mean deviance coefficients among all 72 conditions for two different models is displayed in Table 2 and Table 3. In general, the results in Table 2 and Table 3 reveal a strong association between the sample size and the deviance estimates for these two models. As the sample size increased, the deviance estimates for models also increased. Similar trends were found for the AIC and BIC coefficients as well (see Tables 2 and 3).

Compared with the Rasch testlet model, the Rasch subdimension model always had a smaller deviance, AIC and BIC value under the same condition. As the association between testlets increased, the discrepancy of the model fit indices between the Rasch subdimension model and the Rasch testlet model increased as well. Therefore, according to the model fit results in Table 2 and Table 3, the Rasch subdimension model demonstrated a better performance than the Rasch testlet model when the associations between testlets existed.

<u>Bias and RMSE</u>

In order to reveal how bias and RMSE changes as a function of ability variation, the ability range was split into 6 intervals and the bias and RMSE estimates are calculated accordingly. Table 4 to Table 7 display the mean bias estimates of ability (θ) estimate recovery (i.e. EAP estimate) with 6 different ability intervals for two different models over all 72 conditions. According to the results listed in the tables, a relatively high magnitude of positive bias was observed at the lowest ability interval level ($\theta \leq -2.0$) for both models across all conditions. Meanwhile, relatively high magnitude of negative bias was also found at the highest ability interval level ($\theta \leq -2.0$) for both models across all conditions. Since applying EAP estimation might result in the ability estimate distribution leaning towards its mean, a possible cause for this high magnitude of bias at both ends of the ability intervals might be the usage of the EAP estimates. Other than that high magnitude of bias at both ends of the ability interval phenomena, no obvious patterns and associations between mean bias variations and the major factors in this study were found across both the Rasch testlet model and the Rasch subdimension model.

In addition, Table 8 to Table 11 display the RMSE estimates of ability (θ) estimate recovery with 6 different ability intervals for two different models over all 72 conditions. Similar to the bias estimates, except for that relatively high magnitude of RMSE estimates at both ends of the ability intervals, no obvious patterns and associations between RMSE estimate variations and the major factors in this study were found across two models either.

In sum, both models performed fairly well in ability estimates recovery on the basis of the relatively low magnitude of bias and RMSE estimates from the analysis results.

Test Reliability

A summary of the test reliability analyses is presented in Tables 12 and 13. Two columns of estimates were provided for each model of each condition. For most of the conditions, the reliability estimates from the Rasch testlet model were higher than the reliability estimates from the Rasch subdimension model. The association between test reliability and other factors are described below.

First, the difference in test reliability estimates between the Rasch testlet model and the Rasch subdimension model indicated an association between the magnitude of the correlation between testlets and the test reliability overestimation. In general, the magnitude of the test reliability analyzed from the Rasch testlet model is slightly higher than its corresponding coefficient from the Rasch subdimension model (within 0.02). As the magnitude of the correlation between testlets increased (i.e. from 0.1 to 0.3), the extent of test reliability overestimation for the Rasch testlet model is supposed to increase as well. However, no obvious patterns and associations were found between the magnitude variation of the correlation between testlets and the test reliability overestimation for the Rasch testlet model. This phenomenon occurred because of the small magnitude of the between testlets association (i.e. 0.1-0.3) chosen in this study. Second, as we mentioned before, the ratio of the correlated/total testlet numbers is very important in research. However, for simplicity of the study, only three correlated testlets were included. Theoretically, only one pair of correlated testlets was included. So, no variations of the ratio of the correlated/total testlet numbers exist in this study. Finally, no evident patterns were found to disclose the association between test reliability and testlet variance.

AN EMPIRICAL CASE

The National Board of Osteopathic of Medical Examiners (NBOME) offers computer-based COMLEX-USA exams online. This computer-based exam series is designed to assess the osteopathic medical knowledge and clinical skills considered essential for osteopathic generalist physicians to practice medicine without supervision. The COMLEX-USA exam responses have been analyzed with the standard Rasch IRT Model. The 2008 National Board of Osteopathic of Medical Examiners (NBOME) COMLEX-USA Level-2 exam data was used as an empirical case for this study. The COMLEX-USA level-2 exam consisted of 350 items in 7 blocks including 141 independent items and 209 testlet items grouped in 95 testlets (all medium testlet sizes). The item type was identified (i.e. A -single item, D-single Item with graph, B-matching item, S-testlet item, F-testlet item with graph). The B, S, and F type items were categorized as testlet items. Among all 95 testlets, there were 4 testlets with matching items and 9 testlets with a graph. The testlet sizes range from 2 to 4. A total of 450 examinees were included in the examinee population. No missing data existed. The data of the first block of this exam (Block-1) was used for this study. Block-1 data contained 50 items including 27 independent items and 23 testlet items within 10 testlets.

The data set was analyzed using the Rasch testlet model and the Rasch subdimension model separately. The values of deviance for these two models were 19,237.40 and 19,190.02, respectively. The values of AIC for these two models were 19357.40 and 19,310.02, respectively. The values of BIC of these two models were 19970.51 and 19923.13, respectively. The total numbers of estimated parameters for these two models are 60 and 95. According to the model's deviance results, the Rasch subdimension model had a better model fit than the Rasch testlet model, as was the case in our simulation studies. Also, the Rasch subdimension model outperformed the Rasch testlet model, according to their AIC and BIC results.

Furthermore, the estimates of test reliability for the overarching latent trait are 0.891 for the Rasch testlet model, 0.882 for the Rasch subdimension model. Thus, the Rasch testlet model appeared to slightly overestimate the test reliability due to its ignorance of the association between testlets.

In summary, the Rasch subdimension model has a better fit, compared with the Rasch testlet model when used to analyze NBOME COMLEX exams. In addition, the test reliability discrepancy between the Rasch subdimension model and the Rasch testlet model to analyze NBOME COMLEX data is within the range of 0.01. This result also supports the conclusion that the Rasch subdimension model is the better model choice for analyzing NBOME COMLEX exams.

DISCUSSION AND CONCLUSION

In accordance with the simulation results and the empirical case results, several empirical findings related to testlet modeling emerged in this study. First, our results suggest that the Rasch subdimension model performed better than the Rasch testlet model under small testlet sizes and when associations between testlets exist. The results also showed that sample size had a observable effect on the analysis results for the two models. As the sample size increased, the discrepancies between model estimates and the real data set increased. Also, the degree of the test reliability overestimation for the Rasch testlet model slightly increased when the sample size increased.

Second, the bias and RMSE results from the process of the ability parameter recovery indicated that no evident pattern can be found to reveal the association between the factor variations (i.e., the sample size, the association between testlets) and the bias/RMSE result. The magnitude of the testlet variance did not have an evident impact on the accuracy of the ability estimation. However, this study only investigates a small range of the testlet variance (i.e. [0,1]). A broader range of the testlet variance is worthy of more investigation. Using EAP estimates has major effects on the bias and RMSE results; they each change at both tails of ability distribution. In sum, because these two models are both Rasch type models, the precision of the ability parameter recovery for these models is relatively good. Both Rasch type models do show robustness, to some extent, when handling associations between testlets.

Although there was no obvious discrepancy of the test reliability estimates between the Rasch testlet model and the Rasch subdimension model, a small overestimation trend merged from the Rasch testlet model test reliability estimation. We found no distinguishable difference of the test reliability estimation between the Rasch testlet model and the Rasch subdimension model. We offer two explanations for this lack of finding. First, only small magnitudes of between testlet correlations were chosen for this study (i.e. 0.1, 0.2, and 0.3). Second, for simplicity, only three correlated testlets were included in this study. Therefore, as the testlet number increased per condition in the study, the ratio of the correlated/total testlet numbers decreased. Thus, investigations of larger magnitudes of between testlet correlations and changes in the number of correlated testlets pairs are needed. Because of limited time, we do not explore these issues further in this study. For future research, it is worthwhile to include these two factor variations in situations.

This study compares the performance of two different models in small testlet size situations across changes in sample size, variation of the testlet variance, and the changes of the association between some testlets. The study findings indicate that the Rasch testlet model is still robust as long as the associations between testlets and the pairs of the correlated testlets remain small. Although, under this small between-testlet association situation, the Rasch testlet model shows some robustness, Rasch subdimension model does display a better performance than the Rasch testlet model.

The investigation of the models used to analyze the testlet items based on the between-testlet association circumstances, provides guidance for model selection for future testlet-type data analysis. The Rasch subdimension model offers an advantage over the Rasch testlet model as it allows the association between testlets and better ability parameter estimations when the covariates between testlets exist.

References:

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

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

Brandt, S. (2007a). *Applications of a Rasch model with subdimensions.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.

Brandt, S. (2007b). *Item bundles with items relating to different subtests and their influence on subtests' measurement characteristics.* Paper presented at the 2007 Annual Conference of the American Educational Research Association (AERA), Chicago.

Brandt, S. (2008). *Modeling tests with subtests* (Paper submitted for publication).

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

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

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical response models. *Journal of Educational Measurement*, *26*, 247-260.

Wainer, H. & Kiely, G, L. (1987). Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement, 24, 185-201.

Wainer. H., Lewis. C. (1990). Toward a Psychometrics for Testlets. *Journal of Educational Measurement. 27(1),* 1-14.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 LawSchool Admissions Test as an example. *Applied Measurement in Education*, *8*, 157-186.

Wang, W.-C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological*

Measurement, *29*(2), 126–149.

Zhang, O., Shen, L., Cannady, M. (2009). *Polytomous IRT or Testlet Model: An Evaluation of Scoring Models under Small Testlet Size Situation"*.Paper presented at The 15th International Objective Measurement Workshop (IOMW 2010), Boulder.

Appendix: TABLE

Table 1 Study Design Condition

Condition	sample size	Testlet association	Testlet Number	Testlet Variance	Condition	Sample size	Testlet associatio n	Testlet Number	Testlet Variance
1	1000	0.1	9	0.25	37	500	0.1	9	0.25
2				0.5	38				0.5
3				0.75	39				0.75
4				1	40				1
5			6	0.25	41			6	0.25
6				0.5	42				0.5
7				0.75	43				0.75
8				1	44				1
9			3	0.25	45			3	0.25
10				0.5	46				0.5
11				0.75	47				0.75
12				1	48				1
13		0.2	9	0.25	49		0.2	9	0.25
14				0.5	50				0.5
15				0.75	51				0.75
16				1	52				1
17			6	0.25	53			6	0.25
18				0.5	54				0.5
19				0.75	55				0.75
20				1	56				1
21			3	0.25	57			3	0.25
22				0.5	58				0.5
23				0.75	59				0.75
24				1	60				1
25		0.3	9	0.25	61		0.3	9	0.25
26				0.5	62				0.5
27				0.75	63				0.75
28				1	64				1
29			6	0.25	65			6	0.25
30				0.5	66				0.5
31				0.75	67				0.75
32				1	68				1
33			3	0.25	69			3	0.25
34				0.5	70				0.5
35				0.75	71				0.75
36				1	72				1

Table 2. Rasch Testlet Model vs Rasch Subdimension Model-Deviance, AIC, BIC (Sample size 1000)

	Rasch Testlet Model				Rasch Subdimension Model			
Condition	No.Parameters	mean.deviance	mean.AIC	mean.BIC	No.Parameters	mean.deviance	mean.AIC	mean.BIC
1	69	71091.4976	71229.4976	72044.7679	96	70944.8651	71136.8651	72271.1541
2	69	75077.7995	75215.7995	76031.0697	96	74944.4365	75136.4365	76270.7255
3	69	75249.1993	75387.1993	76202.4695	96	75063.5283	75255.5283	76389.8173
4	69	74986.5889	75124.5889	75939.8591	96	74833.9646	75025.9646	76160.2536
5	66	69053.3789	69185.3789	69965.2026	75	68910.8756	69060.8756	69947.0389
6	66	70932.7197	71064.7197	71844.5434	75	70871.5578	71021.5578	71907.7211
7	66	78140.6321	78272.6321	79052.4558	75	78011.5280	78161.5280	79047.6913
8	66	76443.1713	76575.1713	77354.9950	75	76362.0526	76512.0526	77398.2159
9	63	74116.0070	74242.0070	74986.3842	63	73992.1193	74118.1193	74862.4965
10	63	78042.0984	78168.0984	78912.4755	63	78001.2896	78127.2896	78871.6668
11	63	72976.9271	73102.9271	73847.3043	63	72910.8105	73036.8105	73781.1876
12	63	76790.1390	76916.1390	77660.5162	63	76696.8423	76822.8423	77567.2195
13	69	70074.4890	70212.4890	71027.7592	96	69914.2401	70106.2401	71240.5291
14	69	70249.3276	70387.3276	71202.5978	96	70047.3362	70239.3362	71373.6253
15	69	75235.1419	75373.1419	76188.4121	96	75098.1554	75290.1554	76424.4444
16	69	76495.1190	76633.1190	77448.3892	96	76342.3972	76534.3972	77668.6862
17	66	72001.4514	72133.4514	72913.2751	75	71867.5920	72017.5920	72903.7553
18	66	72541.5484	72673.5484	73453.3721	75	72421.9628	72571.9628	73458.1261
19	66	74068.5130	74200.5130	74980.3367	75	73965.6535	74115.6535	75001.8168
20	66	77324.8610	77456.8610	78236.6847	75	77160.9941	77310.9941	78197.1574
21	63	72166.4972	72292.4972	73036.8744	63	72042.1848	72168.1848	72912.5619
22	63	75975.0048	76101.0048	76845.3819	63	75858.9916	75984.9916	76729.3688
23	63	74450.6895	74576.6895	75321.0667	63	74376.2149	74502.2149	75246.5921
24	63	76122.1356	76248.1356	76992.5127	63	76075.1884	76201.1884	76945.5655
25	69	72099.6201	72237.6201	73052.8904	96	71960.5665	72152.5665	73286.8555
26	69	71091.7240	71229.7240	72044.9942	96	70882.6840	71074.6840	72208.9730
27	69	74892.3702	75030.3702	75845.6404	96	74742.6033	74934.6033	76068.8923
28	69	76072.7804	76210.7804	77026.0506	96	75860.4226	76052.4226	77186.7116
29	66	68629.0503	68761.0503	69540.8740	75	68492.2570	68642.2570	69528.4203
30	66	76801.3774	76933.3774	77713.2011	75	76674.5828	76824.5828	77710.7461
31	66	75584.4791	75716.4791	76496.3028	75	75449.2741	75599.2741	76485.4374
32	66	77454.3587	77586.3587	78366.1824	75	77361.6240	77511.6240	78397.7873
33	63	73076.3384	73202.3384	73946.7155	63	75309.5762	75435.5762	76179.9534
34	63	74346.0333	74472.0333	75216.4105	63	77949.3777	78075.3777	78819.7549
35	63	75960.0352	76086.0352	76830.4123	63	74515.0729	74641.0729	75385.4500
36	63	75983.1827	76109.1827	76853.5598	63	72604.7487	72730.7487	73475.1259

Table 3. Rasch Testlet Model vs Rasch Subdimension Model-Deviance, AIC, BIC (Sample size 500)

		Rasch Testlet Model				Rasch Subdimension Model		
Condition	No.Parameters	mean.deviance	mean.AIC	mean.BIC	No.Parameters	mean.deviance	mean.AIC	mean.BIC
37	69	35347.6628	35485.6628	36205.2788	96	35227.6834	35419.6834	36420.8882
38	69	34956.8802	35094.8802	35814.4961	96	34866.6836	35058.6836	36059.8884
39	69	37253.0905	37391.0905	38110.7064	96	37118.9164	37310.9164	38312.1212
40	69	37673.2579	37811.2579	38530.8738	96	37516.6650	37708.6650	38709.8698
41	66	36820.9132	36952.9132	37641.2414	75	36767.8991	36917.8991	37700.0903
42	66	37545.9103	37677.9103	38366.2386	75	37461.7410	37611.7410	38393.9322
43	66	36990.6222	37122.6222	37810.9504	75	36909.8133	37059.8133	37842.0045
44	66	38120.3490	38252.3490	38940.6772	75	38036.3585	38186.3585	38968.5497
45	63	37018.6096	37144.6096	37801.6502	63	36994.9900	37120.9900	37778.0306
46	63	38524.0002	38650.0002	39307.0408	63	38451.8359	38577.8359	39234.8765
47	63	37846.3967	37972.3967	38629.4373	63	37819.0588	37945.0588	38602.0994
48	63	37406.6078	37532.6078	38189.6484	63	37374.0845	37500.0845	38157.1251
49	69	34359.6843	34497.6843	35217.3002	96	34274.4155	34466.4155	35467.6202
50	69	37304.7111	37442.7111	38162.3270	96	37232.2322	37424.2322	38425.4369
51	69	36225.4991	36363.4991	37083.1150	96	36061.1287	36253.1287	37254.3334
52	69	36809.7967	36947.7967	37667.4126	96	36701.1341	36893.1341	37894.3389
53	66	35646.8722	35778.8722	36467.2005	75	35558.4937	35708.4937	36490.6850
54	66	36797.1420	36929.1420	37617.4703	75	36690.3689	36840.3689	37622.5601
55	66	37522.3641	37654.3641	38342.6924	75	37450.6857	37600.6857	38382.8769
56	66	37470.5538	37602.5538	38290.8820	75	37380.3303	37530.3303	38312.5215
57	63	37500.6764	37626.6764	38283.7170	63	37439.1707	37565.1707	38222.2113
58	63	36192.3341	36318.3341	36975.3747	63	36189.1720	36315.1720	36972.2127
59	63	36288.9676	36414.9676	37072.0082	63	36249.1522	36375.1522	37032.1928
60	63	37603.6148	37729.6148	38386.6554	63	37569.2266	37695.2266	38352.2673
61	69	34692.2131	34830.2131	35549.8290	96	34606.2033	34798.2033	35799.4081
62	69	35121.1175	35259.1175	35978.7334	96	35038.4441	35230.4441	36231.6489
63	69	38007.6510	38145.6510	38865.2669	96	37841.9821	38033.9821	39035.1869
64	69	37806.8789	37944.8789	38664.4948	96	37594.1003	37786.1003	38787.3051
65	66	34564.1633	34696.1633	35384.4916	75	34497.7797	34647.7797	35429.9709
66	66	36066.0798	36198.0798	36886.4080	75	35968.9090	36118.9090	36901.1003
67	66	37770.4718	37902.4718	38590.8001	75	37688.0447	37838.0447	38620.2359
68	66	38927.6577	39059.6577	39747.9860	75	38865.4961	39015.4961	39797.6873
69	63	38192.8896	38318.8896	38975.9302	63	38143.9643	38269.9643	38927.0049
70	63	38266.3271	38392.3271	39049.3677	63	38239.9349	38365.9349	39022.9755
71	63	38309.6341	38435.6341	39092.6748	63	38261.2428	38387.2428	39044.2835
72	63	38750.6334	38876.6334	39533.6741	63	38723.3559	38849.3559	39506.3965

Table 4. Rasch Testlet Model Bias of ability estimate recovery (EAP)-Sample Size 1000

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$ mean.bias1	$-2.0 < \theta \le -1.0$ mean.bias2	$-1.0 < \theta \le 0.0$ mean.bias3	$0.0 < \theta \le 1.0$ mean.bias4	$1.0 < \theta \le 2.0$ mean.bias5	$\theta > 2.0$ mean.bias6
1	0.1	9	0.25	2.093603	1.3291	0.4137	-0.4950	-1.4460	-2.4942
2			0.5	1.831588	1.3307	0.4256	-0.4638	-1.3697	-2.3617
3			0.75	2.555581	1.4065	0.4532	-0.5184	-1.3843	-2.6258
4			1	2.317813	1.3311	0.4068	-0.4984	-1.3842	-2.3948
5		6	0.25	2.445479	1.4558	0.5063	-0.4077	-1.2975	-2.3594
6			0.5	2.003147	1.4194	0.4907	-0.4251	-1.3937	-2.3259
7			0.75	1.831454	1.4252	0.5533	-0.4015	-1.3170	-2.2927
8			1	2.004335	1.3714	0.5142	-0.4376	-1.3297	-2.1838
9		3	0.25	2.100234	1.4437	0.5276	-0.4350	-1.2840	-2.4427
10			0.5	2.097916	1.4451	0.5105	-0.4404	-1.3541	-2.2087
11			0.75	2.31924	1.3845	0.5107	-0.4482	-1.3580	-2.3240
12			1	2.204876	1.3901	0.5105	-0.4557	-1.3717	-2.2294
13	0.2	9	0.25	2.585524	1.3339	0.4112	-0.4918	-1.4577	-2.4677
14			0.5	1.552057	1.3268	0.4287	-0.4920	-1.4318	-2.3640
15			0.75	2.002734	1.3912	0.4201	-0.5021	-1.3824	-2.5468
16			1	2.3177	1.3143	0.4327	-0.4707	-1.3810	-2.4217
17		6	0.25	1.75489	1.4227	0.5443	-0.3881	-1.3093	-2.3410
18			0.5	2.096949	1.3897	0.4593	-0.3960	-1.3391	-2.3615
19			0.75	2.556602	1.4141	0.4749	-0.4498	-1.3383	-2.2518
20			1	1.682674	1.3937	0.4937	-0.4284	-1.3102	-2.3994
21		3	0.25	2.003535	1.4004	0.4778	-0.4545	-1.3369	-2.4143
22			0.5	2.318934	1.3953	0.4855	-0.4291	-1.3539	-2.3536
23			0.75	2.581009	1.3877	0.4862	-0.4457	-1.3290	-2.2577
24			1	1.552078	1.4282	0.5013	-0.4165	-1.3693	-2.2253
25	0.3	9	0.25	1.830828	1.3859	0.4276	-0.5198	-1.3666	-2.3497
26			0.5	2.581605	1.3122	0.3857	-0.5410	-1.4676	-2.3531
27			0.75	1.912346	1.3614	0.4076	-0.4894	-1.4195	-2.3576
28			1	1.491663	1.3959	0.4656	-0.4444	-1.3636	-2.4268
29		6	0.25	2.317677	1.4115	0.4694	-0.4259	-1.3841	-2.3175
30			0.5	1.615573	1.4374	0.5016	-0.4023	-1.3085	-2.1444
31			0.75	1.493041	1.4384	0.5129	-0.4120	-1.3093	-2.3041
32			1	2.442879	1.4187	0.4996	-0.4363	-1.4045	-2.3886
33		3	0.25	1.493068	1.3977	0.4747	-0.4236	-1.3118	-2.2847
34			0.5	1.615977	1.3571	0.4826	-0.4435	-1.3554	-2.3851
35			0.75	3.558036	1.4485	0.4941	-0.4395	-1.3071	-2.3868
36			0.1	2.444128	1.4901	0.4939	-0.4115	-1.3613	-2.4852

Table 5. Rasch Testlet Model Bias of ability estimate recovery (EAP)-Sample Size 500

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$ mean.bias1	$-2.0 < \theta \le -1.0$ mean.bias2	$-1.0 < \theta \le 0.0$ mean.bias3	$0.0 < \theta \le 1.0$ mean.bias4	$1.0 < \theta \le 2.0$ mean.bias5	$\theta > 2.0$ mean.bias6
37	0.1	9	0.25	2.2917	1.3104	0.3977	-0.5290	-1.4747	-2.3565
38			0.5	2.5565	1.4086	0.3914	-0.5181	-1.4410	-2.2792
39			0.75	2.2930	1.3668	0.4427	-0.4983	-1.4427	-2.3410
40			1	2.8369	1.3371	0.3867	-0.5530	-1.4739	-2.6776
41		6	0.25	2.0938	1.4941	0.5451	-0.3513	-1.3329	-2.2716
42			0.5	2.1684	1.4390	0.4669	-0.4582	-1.3913	-2.1727
43			0.75	2.5542	1.4136	0.4836	-0.4284	-1.3449	-2.3790
44			1	2.1674	1.4490	0.5162	-0.4506	-1.3139	-2.3847
45		3	0.25	2.1674	1.3606	0.4670	-0.5021	-1.3342	-2.3091
46			0.5	2.2907	1.4463	0.4724	-0.4347	-1.3674	-2.2619
47			0.75	2.0176	1.3992	0.4916	-0.4390	-1.4104	-2.2547
48			1	2.5562	1.3495	0.5340	-0.3991	-1.3688	-2.2189
49	0.2	9	0.25	2.1679	1.3333	0.4261	-0.5014	-1.4068	-2.5533
50			0.5	2.8374	1.4021	0.4401	-0.5012	-1.4216	-2.3496
51			0.75	2.1979	1.3541	0.4220	-0.4980	-1.4335	-2.3917
52			1	2.9038	1.2479	0.3937	-0.5258	-1.4349	-2.3305
53		6	0.25	2.7347	1.4413	0.4750	-0.4394	-1.2847	-2.1443
54			0.5	2.0949	1.4363	0.5045	-0.4141	-1.2987	-2.3237
55			0.75	2.1670	1.3721	0.4888	-0.4767	-1.3585	-2.4525
56			1	2.5548	1.4570	0.4996	-0.3787	-1.3736	-2.1215
57		3	0.25	2.0179	1.4308	0.4578	-0.4549	-1.3750	-2.3465
58			0.5	2.1969	1.3446	0.4790	-0.4493	-1.3487	-2.2369
59			0.75	2.5848	1.3865	0.4714	-0.4431	-1.3181	-2.3665
60			1	2.5591	1.4724	0.5090	-0.4514	-1.3423	-2.3671
61	0.3	9	0.25	2.8377	1.3610	0.4017	-0.5203	-1.4100	-2.4297
62			0.5	2.8380	1.3202	0.4002	-0.4695	-1.3757	-2.3418
63			0.75	2.5535	1.3646	0.4662	-0.4529	-1.3487	-2.3192
64			1	2.0177	1.3481	0.3991	-0.5102	-1.5033	-2.4119
65		6	0.25	2.5565	1.4380	0.4880	-0.4350	-1.3248	-2.2350
66			0.5	2.5568	1.4939	0.5403	-0.4040	-1.3523	-2.2629
67			0.75	2.5553	1.4008	0.5363	-0.3464	-1.3359	-2.2304
68			1	2.0964	1.3632	0.5098	-0.4166	-1.3384	-2.4560
69		3	0.25	3.5563	1.4507	0.4989	-0.4485	-1.3758	-2.2306
70			0.5	2.5545	1.4054	0.4369	-0.4339	-1.3854	-2.4035
71			0.75	2.1666	1.4062	0.5347	-0.4082	-1.3518	-2.1915
72			0.1	2.9065	1.3818	0.5479	-0.4341	-1.3869	-2.2924

Table 6. Rasch Subdimension Model Bias of ability estimate recovery (EAP)-Sample Size 1000

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$	$-2.0 < \theta \le -1.0$	$-1.0 < \theta \le 0.0$	$0.0 < \theta \le 1.0$	$1.0 < \theta \le 2.0$	$\theta > 2.0$
				mean.bias1	mean.bias2	mean.bias3	mean.bias4	mean.bias5	mean.bias6
1	0.1	9	0.25	3.0944	1.4077	0.4939	-0.4143	-1.3644	-2.4203
2			0.5	1.8323	1.3982	0.4928	-0.3971	-1.3013	-2.2882
3			0.75	3.5563	1.4822	0.5247	-0.4431	-1.3074	-2.5598
4			1	2.3183	1.3944	0.4742	-0.4326	-1.3187	-2.3316
5		6	0.25	2.4450	1.3993	0.4523	-0.4600	-1.3468	-2.4332
6			0.5	2.0029	1.3792	0.4574	-0.4563	-1.4243	-2.3599
7			0.75	1.8308	1.3625	0.4892	-0.4659	-1.3821	-2.3508
8			1	2.0039	1.3408	0.4863	-0.4672	-1.3530	-2.2107
9		3	0.25	2.1001	1.4278	0.5149	-0.4517	-1.2986	-2.4539
10			0.5	2.0979	1.4255	0.4934	-0.4560	-1.3742	-2.2288
11			0.75	2.3193	1.3782	0.5086	-0.4487	-1.3624	-2.3219
12			1	2.2049	1.3908	0.5166	-0.4474	-1.3671	-2.2173
13	0.2	9	0.25	2.5865	1.4334	0.5075	-0.3922	-1.3617	-2.3774
14			0.5	1.5530	1.4059	0.5097	-0.4074	-1.3532	-2.2741
15			0.75	2.0035	1.4667	0.4923	-0.4322	-1.3127	-2.4803
16			1	2.3184	1.3775	0.4987	-0.4069	-1.3172	-2.3533
17		6	0.25	1.7544	1.3679	0.4887	-0.4434	-1.3691	-2.3922
18			0.5	3.0967	1.3558	0.4349	-0.4254	-1.3693	-2.3949
19			0.75	2.5562	1.3866	0.4513	-0.4714	-1.3628	-2.2827
20			1	1.6821	1.3513	0.4507	-0.4673	-1.3544	-2.4489
21		3	0.25	2.0035	1.3916	0.4690	-0.4634	-1.3505	-2.4120
22			0.5	2.3189	1.3822	0.4726	-0.4450	-1.3636	-2.3591
23			0.75	2.5811	1.3809	0.4790	-0.4549	-1.3342	-2.2630
24			1	1.5520	1.4197	0.4909	-0.4264	-1.3797	-2.2407
25	0.3	9	0.25	1.8316	1.4754	0.5169	-0.4359	-1.2804	-2.2664
26			0.5	2.5827	1.4213	0.4920	-0.4368	-1.3669	-2.2373
27			0.75	1.9131	1.4384	0.4896	-0.4089	-1.3375	-2.2898
28			1	1.4921	1.4432	0.5164	-0.3964	-1.3154	-2.3745
29		6	0.25	2.3175	1.3911	0.4488	-0.4448	-1.3985	-2.3331
30			0.5	1.6149	1.3742	0.4357	-0.4652	-1.3730	-2.1990
31			0.75	1.4923	1.3744	0.4443	-0.4804	-1.3782	-2.3813
32			1	2.4426	1.3963	0.4766	-0.4577	-1.4271	-2.4119
33		3	0.25	3.0946	1.4289	0.5016	-0.4403	-1.3495	-2.2820
34			0.5	1.2415	1.3886	0.4866	-0.4638	-1.3517	-2.3799
35			0.75	1.8314	1.4158	0.4826	-0.4413	-1.3971	-2.3057
36			0.1	2.3185	1.3650	0.4973	-0.4255	-1.3189	-2.2391

Table 7. Rasch Subdimension Model Bias of ability estimate recovery (EAP)-Sample Size 500

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \leq -2.0$	$-2.0 < \theta \leq -1.0$	$-1.0 < \theta \leq 0.0$	$0.0 < \theta \leq 1.0$	$1.0 < \theta \leq 2.0$	$\theta > 2.0$
				mean.bias1	mean.bias2	mean.bias3	mean.bias4	mean.bias5	mean.bias6
37	0.1	9	0.25	2.2928	1.4209	0.5027	-0.4269	-1.3626	-2.2655
38			0.5	2.5573	1.4991	0.4846	-0.4290	-1.3445	-2.1928
39			0.75	2.2934	1.4161	0.4914	-0.4479	-1.3936	-2.3022
40			1	2.8379	1.4406	0.4935	-0.4441	-1.3701	-2.5489
41		6	0.25	2.0933	1.4229	0.4720	-0.4260	-1.4049	-2.3405
42			0.5	2.1681	1.4217	0.4522	-0.4757	-1.4141	-2.1921
43			0.75	2.5537	1.3646	0.4398	-0.4680	-1.3911	-2.4327
44			1	2.1672	1.4143	0.4848	-0.4805	-1.3428	-2.4020
45		3	0.25	2.1674	1.3574	0.4670	-0.5000	-1.3383	-2.3136
46			0.5	2.2908	1.4517	0.4808	-0.4286	-1.3604	-2.2680
47			0.75	2.0175	1.4005	0.4954	-0.4384	-1.4121	-2.2498
48			1	2.5560	1.3208	0.5123	-0.4208	-1.3932	-2.2357
49	0.2	9	0.25	2.1688	1.4279	0.5133	-0.4148	-1.3289	-2.4750
50			0.5	2.8380	1.4698	0.5092	-0.4354	-1.3538	-2.2868
51			0.75	2.1987	1.4415	0.5155	-0.4077	-1.3471	-2.3211
52			1	2.9052	1.3757	0.5260	-0.3945	-1.3052	-2.1788
53		6	0.25	2.7345	1.4175	0.4498	-0.4650	-1.3105	-2.1605
54			0.5	2.0943	1.3817	0.4581	-0.4608	-1.3416	-2.3785
55			0.75	2.1667	1.3438	0.4632	-0.5035	-1.3860	-2.4719
56			1	2.5543	1.3970	0.4489	-0.4310	-1.4270	-2.1910
57		3	0.25	2.0181	1.4343	0.4572	-0.4538	-1.3670	-2.3464
58			0.5	2.1970	1.3370	0.4753	-0.4511	-1.3565	-2.2351
59			0.75	2.5847	1.3741	0.4589	-0.4553	-1.3366	-2.3733
60			1	2.5590	1.4589	0.4986	-0.4590	-1.3560	-2.3766
61	0.3	9	0.25	2.8383	1.4483	0.4865	-0.4419	-1.3290	-2.3503
62			0.5	2.8387	1.3860	0.4652	-0.4030	-1.3105	-2.2698
63			0.75	2.5540	1.4106	0.5104	-0.4123	-1.3003	-2.2572
64			1	2.0190	1.4453	0.5096	-0.4007	-1.4012	-2.3012
65		6	0.25	2.5562	1.4064	0.4538	-0.4663	-1.3639	-2.2686
66			0.5	2.5558	1.4168	0.4656	-0.4808	-1.4229	-2.3480
67			0.75	2.5548	1.3408	0.4723	-0.4107	-1.3966	-2.2897
68			1	2.0961	1.3271	0.4746	-0.4535	-1.3739	-2.4835
69		3	0.25	2.5562	1.4273	0.4772	-0.4702	-1.3968	-2.2397
70			0.5	2.5543	1.3999	0.4309	-0.4424	-1.4084	-2.4089
71			0.75	2.1665	1.3859	0.5262	-0.4245	-1.3730	-2.2165
72			0.1	2.9064	1.3723	0.5334	-0.4438	-1.3939	-2.3081

Table 8. Rasch Testlet Model RMSE of ability estimate recovery (EAP)-Sample Size 1000

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$	$-2.0 < \theta \le -1.0$	$-1.0 < \theta \le 0.0$	$0.0 < \theta \le 1.0$	$1.0 < \theta \le 2.0$	$\theta > 2.0$
				mean.RMSE1	mean.RMSE2	mean.RMSE3	mean.RMSE4	mean.RMSE5	mean.RMSE6
1	0.1	9	0.25	0.5186	0.0952	0.0127	0.0179	0.1158	0.2283
2			0.5	0.4586	0.1064	0.0149	0.0454	0.1110	0.3430
3			0.75	0.7405	0.1242	0.0497	0.0142	0.1163	0.3055
4			1	0.5860	0.0948	0.0225	0.0346	0.1223	0.3952
5		6	0.25	0.5600	0.1192	0.0131	0.0430	0.1072	0.3773
6			0.5	0.6465	0.1123	0.0158	0.0128	0.0856	0.2097
7			0.75	0.4325	0.1588	0.0353	0.0195	0.1183	0.2722
8			1	0.4727	0.0793	0.0401	0.0326	0.1431	0.3782
9		3	0.25	0.3839	0.1157	0.0119	0.0119	0.1430	0.4031
10			0.5	0.4412	0.1486	0.0475	0.0351	0.1187	0.3013
11			0.75	0.5097	0.1336	0.0395	0.0303	0.1014	0.2862
12			1	0.5970	0.1409	0.0459	0.0421	0.1216	0.4531
13	0.2	9	0.25	0.8506	0.1355	0.0096	0.0289	0.1411	0.4246
14			0.5	0.3345	0.1394	0.0137	0.0240	0.1654	0.4161
15			0.75	0.5215	0.1473	0.0472	0.0163	0.0807	0.1615
16			1	0.4769	0.1244	0.0123	0.0433	0.1131	0.2502
17		6	0.25	0.4140	0.1141	0.0290	0.0434	0.1539	0.3470
18			0.5	0.5970	0.1502	0.0175	0.0357	0.0947	0.2740
19			0.75	0.6518	0.1268	0.0509	0.0300	0.1286	0.2551
20			1	0.4221	0.0683	0.0122	0.0123	0.1192	0.2708
21		3	0.25	0.4528	0.0787	0.0440	0.0193	0.0841	0.2751
22			0.5	0.6509	0.1546	0.0132	0.0392	0.1313	0.2731
23			0.75	0.5045	0.1535	0.0187	0.0336	0.1504	0.3951
24			1	0.4033	0.1379	0.0164	0.0090	0.1586	0.3263
25	0.3	9	0.25	0.4312	0.1246	0.0179	0.0596	0.1031	0.1913
26			0.5	0.4565	0.0961	0.0449	0.0286	0.1640	0.4055
27			0.75	0.5158	0.1619	0.0099	0.0296	0.1255	0.2861
28			1	0.3458	0.1140	0.0156	0.0350	0.1994	0.4658
29		6	0.25	0.5135	0.1722	0.0567	0.0138	0.1006	0.2993
30			0.5	0.3737	0.1384	0.0171	0.0142	0.1196	0.3822
31			0.75	0.4082	0.1667	0.0100	0.0124	0.0996	0.2139
32			1	0.3805	0.0916	0.0281	0.0160	0.1314	0.3830
33		3	0.25	0.4336	0.1218	0.0287	0.0326	0.1631	0.3454
34			0.5	0.3842	0.0945	0.0426	0.0295	0.0933	0.1508
35			0.75	0.6066	0.1183	0.0164	0.0346	0.1444	0.3645
36			0.1	0.4589	0.1231	0.0143	0.0112	0.1530	0.3636

Table 9. Rasch Testlet Model RMSE of ability estimate recovery (EAP)-Sample Size 500

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$	$-2.0 < \theta \le -1.0$	$-1.0 < \theta \le 0.0$	$0.0 < \theta \le 1.0$	$1.0 < \theta \le 2.0$	$\theta > 2.0$
				mean.RMSE1	mean.RMSE2	mean.RMSE3	mean.RMSE4	mean.RMSE5	mean.RMSE6
37	0.1	9	0.25	0.7726	0.1916	0.0444	0.0477	0.2173	0.4432
38			0.5	0.6487	0.2318	0.0305	0.0155	0.2402	0.6794
39			0.75	0.8714	0.1226	0.0155	0.0168	0.2354	0.6060
40			1	0.5320	0.2131	0.0326	0.0880	0.2596	0.6171
41		6	0.25	1.0037	0.1412	0.0618	0.0189	0.1448	0.3833
42			0.5	0.7265	0.2087	0.0438	0.0278	0.2135	0.9089
43			0.75	0.6267	0.1794	0.0355	0.0153	0.1384	0.2735
44			1	0.6862	0.1623	0.0613	0.0655	0.2118	0.5177
45		3	0.25	0.8368	0.1849	0.0289	0.0608	0.1333	0.4667
46			0.5	1.1181	0.1504	0.0194	0.0310	0.2411	1.0069
47			0.75	0.8163	0.2491	0.0319	0.0258	0.1976	0.4496
48			1	0.6738	0.1152	0.0145	0.0833	0.1759	0.3143
49	0.2	9	0.25	0.6097	0.2354	0.0155	0.0468	0.2104	0.4297
50			0.5	0.6208	0.1640	0.0168	0.0403	0.1592	0.4181
51			0.75	0.8468	0.1487	0.0859	0.0482	0.1443	0.2868
52			1	0.6691	0.0980	0.0493	0.0163	0.2118	0.5529
53		6	0.25	0.5504	0.1411	0.0265	0.0663	0.1607	0.4738
54			0.5	0.6226	0.1775	0.0220	0.0185	0.1364	0.4777
55			0.75	1.0201	0.1416	0.0520	0.0614	0.1403	0.3762
56			1	0.8151	0.1954	0.0847	0.0278	0.2345	0.6223
57		3	0.25	0.6827	0.2123	0.0470	0.0150	0.1737	0.4455
58			0.5	0.8323	0.1317	0.0448	0.0356	0.1526	0.4260
59			0.75	0.8394	0.2645	0.0465	0.0463	0.1815	0.5293
60			1	0.9401	0.0989	0.0301	0.0368	0.1236	0.2204
61	0.3	9	0.25	0.5879	0.1467	0.0138	0.0195	0.1376	0.2839
62			0.5	0.5192	0.1590	0.0131	0.0185	0.1810	0.5209
63			0.75	0.7272	0.1772	0.0628	0.0382	0.1716	0.4367
64			1	0.7564	0.1218	0.0143	0.0212	0.2054	0.3985
65		6	0.25	0.6816	0.1932	0.0663	0.0743	0.1172	0.4555
66			0.5	0.8144	0.2508	0.0449	0.0141	0.1594	0.4871
67			0.75	0.6944	0.1988	0.0730	0.0226	0.1987	0.4796
68			1	1.2064	0.2024	0.0456	0.0692	0.1154	0.3783
69		3	0.25	0.5742	0.1998	0.0532	0.0279	0.1525	0.2432
70			0.5	0.6731	0.1735	0.0169	0.0561	0.1623	0.3624
71			0.75	0.6499	0.1375	0.0536	0.0569	0.1346	0.5552
72			0.1	0.8758	0.1416	0.0608	0.0256	0.1422	0.2573

Table 10. Rasch Subdimension Model RMSE of ability estimate recovery (EAP)-Sample Size 1000

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$	$-2.0 < \theta \le -1.0$	$-1.0 < \theta \le 0.0$	$0.0 < \theta \le 1.0$	$1.0 < \theta \le 2.0$	$\theta > 2.0$
				mean.RMSE1	mean.RMSE2	mean.RMSE3	mean.RMSE4	mean.RMSE5	mean.RMSE6
1	0.1	9	0.25	0.5344	0.1036	0.0137	0.0166	0.1117	0.2295
2			0.5	0.4632	0.1095	0.0126	0.0432	0.1054	0.3321
3			0.75	0.7568	0.1294	0.0531	0.0117	0.1097	0.3450
4			1	0.6008	0.0999	0.0286	0.0330	0.1186	0.2900
5		6	0.25	0.5569	0.1142	0.0118	0.0456	0.1089	0.2055
6			0.5	0.6354	0.1066	0.0157	0.0123	0.0890	0.2700
7			0.75	0.4254	0.1504	0.0282	0.0227	0.1290	0.4227
8			1	0.4742	0.0754	0.0361	0.0335	0.1528	0.4733
9		3	0.25	0.3776	0.1111	0.0124	0.0137	0.1437	0.2128
10			0.5	0.4333	0.1427	0.0460	0.0355	0.1195	0.2972
11			0.75	0.5076	0.1316	0.0371	0.0332	0.1087	0.2586
12			1	0.6078	0.1446	0.0479	0.0432	0.1202	0.3809
13	0.2	9	0.25	0.8750	0.1470	0.0104	0.0248	0.1311	0.3326
14			0.5	0.3539	0.1485	0.0152	0.0195	0.1615	0.3136
15			0.75	0.5392	0.1516	0.0507	0.0143	0.0745	0.2698
16			1	0.4990	0.1283	0.0112	0.0396	0.1061	0.2018
17		6	0.25	0.4135	0.1079	0.0248	0.0446	0.1648	0.3567
18			0.5	0.5900	0.1476	0.0172	0.0370	0.0980	0.1842
19			0.75	0.6534	0.1310	0.0503	0.0336	0.1311	0.2507
20			1	0.4014	0.0688	0.0128	0.0133	0.1246	0.2138
21		3	0.25	0.4554	0.0778	0.0466	0.0201	0.0902	0.2568
22			0.5	0.6386	0.1545	0.0139	0.0411	0.1306	0.2741
23			0.75	0.5024	0.1559	0.0171	0.0327	0.1527	0.4137
24			1	0.3866	0.1368	0.0149	0.0109	0.1568	0.3354
25	0.3	9	0.25	0.4530	0.1287	0.0205	0.0553	0.0991	0.1890
26			0.5	0.4849	0.1038	0.0518	0.0218	0.1545	0.4155
27			0.75	0.5240	0.1674	0.0103	0.0247	0.1235	0.2972
28			1	0.3579	0.1157	0.0177	0.0339	0.1917	0.5023
29		6	0.25	0.5103	0.1655	0.0546	0.0114	0.1085	0.4494
30			0.5	0.3721	0.1331	0.0151	0.0148	0.1271	0.4163
31			0.75	0.3924	0.1614	0.0097	0.0171	0.1029	0.3013
32			1	0.3771	0.0852	0.0262	0.0177	0.1343	0.4470
33		3	0.25	0.6831	0.0907	0.0388	0.0375	0.1165	0.3160
34			0.5	0.5184	0.0973	0.0499	0.0096	0.1527	0.5366
35			0.75	0.4823	0.0960	0.0115	0.0107	0.0917	0.2239
36			0.1	0.3944	0.1258	0.0115	0.0155	0.0927	0.3120

Table 11. Rasch Subdimension Model RMSE of ability estimate recovery (EAP)-Sample Size 500

condition	Testlet association	Testlet No.	Testlet Variance	$\theta \le -2.0$	$-2.0 < \theta \le -1.0$	$-1.0 < \theta \le 0.0$	$0.0 < \theta \le 1.0$	$1.0 < \theta \le 2.0$	$\theta > 2.0$
				mean.RMSE1	mean.RMSE2	mean.RMSE3	mean.RMSE4	mean.RMSE5	mean.RMSE6
37	0.1	9	0.25	0.8167	0.2069	0.0521	0.0406	0.2071	0.4725
38			0.5	0.6606	0.2381	0.0383	0.0195	0.2248	0.5709
39			0.75	0.8877	0.1318	0.0158	0.0155	0.2308	0.5651
40			1	0.5573	0.2291	0.0405	0.0833	0.2533	0.6207
41		6	0.25	0.9788	0.1365	0.0597	0.0221	0.1537	0.4102
42			0.5	0.6984	0.2094	0.0404	0.0282	0.2114	0.6846
43			0.75	0.6224	0.1654	0.0320	0.0156	0.1456	0.2864
44			1	0.6685	0.1593	0.0621	0.0680	0.2155	0.5846
45		3	0.25	0.8474	0.1857	0.0280	0.0611	0.1321	0.4152
46			0.5	1.1257	0.1473	0.0215	0.0355	0.2349	0.9733
47			0.75	0.8045	0.2517	0.0281	0.0233	0.1987	0.3867
48			1	0.6739	0.1156	0.0128	0.0873	0.1747	0.3510
49	0.2	9	0.25	0.6414	0.2533	0.0142	0.0420	0.2015	0.4880
50			0.5	0.6396	0.1714	0.0203	0.0349	0.1518	0.4112
51			0.75	0.8698	0.1619	0.0925	0.0403	0.1399	0.2543
52			1	0.6910	0.1328	0.0578	0.0175	0.1864	0.5163
53		6	0.25	0.5492	0.1357	0.0258	0.0667	0.1648	0.4780
54			0.5	0.6215	0.1740	0.0174	0.0196	0.1355	0.4891
55			0.75	1.0120	0.1377	0.0509	0.0630	0.1433	0.4063
56			1	0.8029	0.1953	0.0809	0.0305	0.2426	0.6409
57		3	0.25	0.6946	0.2191	0.0459	0.0150	0.1705	0.3765
58			0.5	0.8297	0.1325	0.0437	0.0305	0.1547	0.4820
59			0.75	0.8369	0.2658	0.0417	0.0471	0.1901	0.6770
60			1	0.9264	0.1017	0.0305	0.0356	0.1226	0.3160
61	0.3	9	0.25	0.6057	0.1582	0.0172	0.0196	0.1314	0.2463
62			0.5	0.5450	0.1688	0.0157	0.0213	0.1697	0.4404
63			0.75	0.7446	0.1819	0.0640	0.0363	0.1660	0.4840
64			1	0.7986	0.1346	0.0142	0.0182	0.1967	0.2896
65		6	0.25	0.6786	0.1939	0.0624	0.0743	0.1240	0.2882
66			0.5	0.7781	0.2450	0.0435	0.0193	0.1779	0.5864
67			0.75	0.6648	0.1864	0.0658	0.0267	0.2037	0.4478
68			1	1.2054	0.1974	0.0429	0.0707	0.1188	0.2704
69		3	0.25	0.5584	0.1936	0.0465	0.0272	0.1533	0.2974
70			0.5	0.6827	0.1679	0.0182	0.0565	0.1683	0.3128
71			0.75	0.6510	0.1333	0.0524	0.0573	0.1384	0.5632
72			0.1	0.8825	0.1422	0.0539	0.0294	0.1406	0.2673

Table 12. Test Reliability (Rasch Testlet Model vs. Rasch Subdimension Model)-Sample Size 1000

Sample size 1000					
Condition	Testlet Association	Testlet No.	Testlet Variance	Rasch Testlet model	Rasch Subdimension Model
1	0.1	9	0.25	0.9058	0.9024
2			0.5	0.8946	0.8913
3			0.75	0.9131	0.9119
4			1	0.8662	0.8612
5		6	0.25	0.9085	0.9032
6			0.5	0.9119	0.9050
7			0.75	0.8722	0.8666
8			1	0.8997	0.8860
9		3	0.25	0.8884	0.8735
10			0.5	0.8637	0.8562
11			0.75	0.8921	0.8849
12			1	0.8650	0.8555
13	0.2	9	0.25	0.9084	0.9048
14			0.5	0.9004	0.8965
15			0.75	0.8977	0.8866
16			1	0.9035	0.8974
17		6	0.25	0.8996	0.8855
18			0.5	0.9048	0.9104
19			0.75	0.9152	0.9005
20			1	0.9065	0.9027
21		3	0.25	0.8886	0.8809
22			0.5	0.8809	0.8727
23			0.75	0.8878	0.8802
24			1	0.8851	0.8785
25	0.3	9	0.25	0.9150	0.9011
26			0.5	0.8913	0.8848
27			0.75	0.8951	0.8900
28			1	0.8931	0.8891
29		6	0.25	0.9163	0.9008
30			0.5	0.8842	0.8718
31			0.75	0.8771	0.8573
32			1	0.9093	0.8992
33		3	0.25	0.8897	0.8748
34			0.5	0.8815	0.8731
35			0.75	0.8670	0.8630
36			1	0.8970	0.8809

Table 13. Test Reliability (Rasch Testlet Model vs. Rasch Subdimension Model)-Sample Size 500

Sample size 500					
Condition	Testlet Association	Testlet No.	Testlet Variance	Rasch Testlet model	Rasch Subdimension Model
37	0.1	9	0.25	0.9035	0.9019
38			0.5	0.9139	0.9068
39			0.75	0.8938	0.8841
40			1	0.8802	0.8737
41		6	0.25	0.8952	0.8828
42			0.5	0.9006	0.8970
43			0.75	0.8955	0.8815
44			1	0.8707	0.8771
45		3	0.25	0.8837	0.8742
46			0.5	0.8649	0.8546
47			0.75	0.8792	0.8721
48			1	0.8836	0.8769
49	0.2	9	0.25	0.8948	0.8912
50			0.5	0.8629	0.8553
51			0.75	0.9049	0.9003
52			1	0.9181	0.9077
53		6	0.25	0.9005	0.8966
54			0.5	0.8910	0.8873
55			0.75	0.8947	0.9023
56			1	0.8539	0.8527
57		3	0.25	0.8730	0.8651
58			0.5	0.8788	0.8730
59			0.75	0.8905	0.8859
60			1	0.8812	0.8757
61	0.3	9	0.25	0.9029	0.8982
62			0.5	0.8976	0.8961
63			0.75	0.8880	0.8836
64			1	0.8861	0.8842
65		6	0.25	0.8951	0.8891
66			0.5	0.8834	0.8788
67			0.75	0.8680	0.8677
68			1	0.8665	0.8618
69		3	0.25	0.9130	0.9104
70			0.5	0.8813	0.8743
71			0.75	0.8739	0.8653
72			1	0.8791	0.8721

Table 14 COMLEX-Level 2 2008 Block-1 Item sequence and unit ID

item seq	CASE Type	UNIT_TYPE
1	independent	A
2	independent	A
3	independent	A
4	independent	A
5	independent	A
6	independent	A
7	independent	A
8	independent	A
9	independent	A
10	independent	A
11	independent	A
12	independent	A
13	independent	A
14	independent	A
15	independent	A
16	independent	A
17	independent	A
18	independent	A
19	independent	A
20	independent	A
21	independent	A
22	independent	A
23	independent	A
24	independent	A
25	independent	A
26	independent	D
27	independent	D
28	testlet1	B
29	testlet1	B
30	testlet1	B
31	testlet1	B
32	testlet2	F
33	testlet2	F
34	testlet3	S
35	testlet3	S
36	testlet3	S
37	testlet4	S
38	testlet4	S
39	testlet5	S
40	testlet5	S
41	testlet6	S
42	testlet6	S
43	testlet7	S
44	testlet7	S
45	testlet8	S
46	testlet8	S
47	testlet9	S
48	testlet9	S
49	testlet10	S
50	testlet10	S

Note: Testlet item 28-50; Testlets number: 10; Independent items: 1-27