JOURNAL OF APPLIED MEASUREMENT, *10*(4), 394-407

Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model

Insu Paek*
Harcourt Assessment

Haniza Yon
Educational Testing Service

Mark Wilson
University of California at Berkeley

Taehoon Kang
University of California at Los Angeles

The current Rasch testlet model (RT) assumes independence of the testlet effect and the target dimension. This article investigated the impact of the violation of that assumption on RT and the performance of an extended Rasch testlet model (ET) in which the random parameter variance-covariance matrix is estimated without any constraints. Our simulation results showed that ET was the same or superior to RT in its performance. The target dimension variance in RT was the most strongly affected parameter and the bias of the target dimension variance was largest when the testlet effect was large and the correlation between the testlet effect and the target dimension was high. This suggests that in some real data applications, it may be difficult to accurately assess the size of testlet effect relative to the target dimension. RT showed close performance to ET with regard to item and testlet effect parameter estimation.

* Insu Paek is currently working at Educational Testing Service.

Requests for reprints should be sent to Insu Paek, Educational Testing Service, Rosedale Road, MS – 02P, Princeton, NJ 08541, USA, e-mail: ipaek@ets.org.

Popular IRT models for dichotomous item response data include the Rasch model (1980), and the 2-parameter and 3-parameter logistic models (Birnbaum, 1968). These models have been widely applied to educational achievement tests, and have been extended to cover a wider range of test designs. Variations that have been developed include polytomous IRT models—rating scale (Andrich, 1978), partial credit (Master, 1982), generalized partial credit (Muraki, 1992), and graded response (Samejima, 1969)—and the facet model (Linacre, 1989) which is typically used to study rater effects. In recent years, models have evolved to accommodate tests that have a testlet structure where sets of items share common stimuli. Wainer and Kiely (1987) referred to these item sets as testlets, although they have sometimes been called item bundles (Rosenbaum, 1988; Wilson and Adams, 1995).

The existence of testlets in a test can reduce the accuracy of estimation of model parameters due to the violation of the common assumption of local independence among items in IRT modeling (see, e.g., Wainer and Thissen, 1996; Wainer and Wang, 2000). A simple way to cope with a test having testlets is to use a polytomous IRT model, treating a testlet as a polytomous item. Another way is to directly incorporate the testlet structure into the IRT model, thereby preserving exact item response patterns. This is not possible in the polytomous approach. Direct testlet IRT modeling uses an additional random effect parameter, which can be thought of as the interaction between an examinee and the testlet. The 3-parameter logistic testlet model for dichotomous response data (see, Bradlow, Wainer, and Wang, 1999; Du, 1998; Wainer, Bradlow, and Du, 2000, Wang, Bradlow, and Wainer, 2002) has the following form:

$$P_{ni}(1) = c_i + (1-c_i)\frac{\exp[a_i(\theta_n - b_i + \gamma_{nd(i)})]}{1+\exp[a_i(\theta_n - b_i + \gamma_{nd(i)})]}, \quad (1)$$

where $P_{ni}(1)$ is the probability of response 1 (correct) for person n on the i^{th} item, a_i is item discrimination, b_i is item difficulty, c_i is a guessing parameter, θ_n is person ability, and $\gamma_{nd(i)}$ is testlet effect, representing the interaction between person n and the i^{th} item nested within the testlet d. This 3-parameter logistic testlet model was usually treated as a parametric Bayesian model, and parameters were usually estimated using the Markov chain Monte Carlo (MCMC) technique due to the complexity of the model. When c_i is equal to zero and a_i is equal to one, Equation 1 reduces to the Rasch testlet model:

$$P_{ni}(1) = \frac{\exp[\theta_n - b_i + \gamma_{nd(i)}]}{1+\exp[\theta_n - b_i + \gamma_{nd(i)}]}. \quad (2)$$

Wang and Wilson (2005) applied the Rasch testlet model in a non-Bayesian context. They showed that the Rasch testlet model is a submodel of the multidimensional random coefficient multinomial logit model (MRCMLM; Adams, Wilson, and Wang, 1997). It was further shown that the use of the existing computer program ConQuest (Wu, Adams, and Wilson, 1998), which has multidimensional modeling capacity, obviated the need for derivation of additional model parameter estimation procedures and programming. Compared to the Bayesian testlet model which needs distributional assumptions for all the model parameters, the Rasch testlet model with ConQuest requires distributional assumptions only for person ability and the testlet effect.

In the context of Equations 1 and 2, the typical distributional assumptions for the person and testlet random parameters are:

$$\theta_n \sim N\left(\mu_\theta, \sigma_\theta^2\right), \quad (3)$$

$$\gamma_{nd(i)} \sim N\left(\mu_{\gamma d}, \sigma_{\gamma d}^2\right), \quad (4)$$

i.e., θ and γ are independently and normally distributed (and the testlet effect parameter of σ_γ is constant for person n within testlet d). The assumption of independent normal distributions for θ and γ arises from considerations of the model parameter estimation and convenience of statistical modeling, rather than from realty-based argument or evidence. It would be more natural to have covariances among θ and γ values, for instance by modeling covariance terms in a multivariate normal distribution:

$$\left(\theta, \boldsymbol{\gamma}'\right) \sim MVN\left(\boldsymbol{\mu}, \Sigma\right), \quad (5)$$

where θ is a scalar, $\boldsymbol{\gamma}$ is a vector having k testlet random effect parameters, μ is a $1 \times (k+1)$ vector, and Σ is a $(k+1) \times (k+1)$ variance-covariance matrix. Fixing all off-diagonal elements in Σ equal to zero reduces to the above regular independent normal testlet model (e.g., for the Rasch testlet model, Equations 2, 3, and 4).

In this article, the regular Rasch testlet model is challenged by the violation of its assumption that the parameters θ and γ and independent. The regular Rasch testlet model is compared with an extended Rasch testlet model (Equations 2 and 5) in which the strict independence between θ and γ is relaxed so that the covariance between these parameters is free to vary. The effect of non-zero covariance in the random parameters θ and γ is investigated through a simulation study on the target dimension variance, the testlet effect, and the item parameters. In addition, applying the likelihood ratio test to the regular and extended Rasch testlet model, statistical model comparisons were conducted for zero-covariance and non-zero covariance cases.

Method

Model

The regular Rasch testlet model and the extended model suggested here share the same form of the item response function (IRF). The difference between them lies in the random parameter variance-covariance structure. The regular Rasch testlet model for dichotomous response data is Equation 2, with the distributional assumptions of Equations 3 and 4. The extended Rasch testlet model for dichotomous response data is Equation 2, with the distributional assumptions of Equation 5. Note that the regular Rasch testlet model is nested within the extended Rasch testlet model. Also, with some algebra and matrix transformation, one can show that both the regular and extended Rasch testlet models are the submodels of the MRCMLM (see also Wang and Wilson, 2005). Wang and Wilson (2005) used the program ConQuest for estimation of the parameters of the regular Rasch testlet model. ConQuest can be used to estimate the parameters of the extended Rasch testlet model, with no need for the development of additional parameter estimation procedures. In contrast to the Bayesian approach, there is no specification of hyperparameters in the extended Rasch testlet model, and all parameters except θ and γ are fixed unknown parameters.

For the model identification of the 3-parameter logistic testlet model, typically

$$\theta_n \sim N(0, 1) \text{ and} \tag{6}$$

$$\gamma_{nd(i)} \sim N\left(0, \sigma^2_{\gamma d}\right) \tag{7}$$

are used. For the extended Rasch model identification,

$$\left(\theta, \boldsymbol{\gamma}'\right) \sim MVN\left(\mathbf{0}, \Sigma\right), \tag{8}$$

is used, where $\mathbf{0}$ is a $1 \times (k+1)$ zero vector. Note that when a test is composed entirely of testlets without independent items, Equation 8 may not be sufficient for the extended testlet model identification. In this study, the extended testlet model with Equation 8 is applied to a test having both independent items and testlets.

Simulation Design and Data Generation

As a first step, a simple scenario was constructed to make a clear inference on the impact of existing covariance structure on the regular testlet model. In other words, the scenario was formulated to demonstrate the effect of dimension correlation in the bivariate normal distribution of θ and γ on the parameters of the Rasch testlet model, especially the target dimension variance σ^2_{θ} and the testlet effect parameter $\sigma^2_{\gamma_d}$. A simulated test had 20 items. As one of the simple cases, the first 10 items were independent items (so that regular IRT model estimation could be legitimately applied) and the last 10 items were supposed to share a common stimulus. Therefore, Equation 8 is a bivariate normal distribution with $\Sigma_{(2x2)}$. By definition, no testlet can contain fewer than 2 items, and 10 is probably close to the upper bound for the number of items in a testlet. Wang and Wilson (2005) used testlets with 5 and 10 items in their simulation studies. The

sample size for each simulation condition was fixed as 1000.

The control variables in the simulation were the size of the testlet effect $\sigma^2_{\gamma_d}$ and the correlation (or covariance) between θ and γ ($\rho_{(\theta,\gamma)}$). The size of the testlet effect had 4 levels: 0.25, 0.5, 0.75, and 1. The target dimension variance σ^2_θ was set to unity under all conditions. Therefore, $\sigma^2_{\gamma_d}$ equal to 0.25 (i.e., 25 % of the target dimension variance) may be considered as a small testlet effect; 0.5 (50% of the target dimension variance) as a medium testlet effect; and 0.75 and 1 (75 and 100% of the target dimension variance) as large testlet effects. These values correspond to those used by Wang and Wilson (2005). For the dimensional correlation, $\rho_{(\theta,\gamma)}$ had 4 levels: 0, 0.2, 0.5, and 0.8. $\rho_{(\theta,\gamma)}$ equal to zero represents independence between θ and γ and is the assumption for the regular testlet model. We used a crossed simulation design, resulting in a total of 16 conditions (4 levels of $\sigma^2_{\gamma_d}$ times 4 levels of $\rho_{(\theta,\gamma)}$). Table 1 summarizes the simulation control variables. Item parameter values were fixed across all conditions. They were randomly drawn from $N(0,1)$. The item parameter values ranged from –1.845 to 1.916 with a mean of –0.016 and a standard deviation of 0.869.

Table 1

Simulated Random Parameter Structure

σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$
1.00	0.00	0.25	0.0
1.00	0.10	0.25	0.2
1.00	0.25	0.25	0.5
1.00	0.40	0.25	0.8
1.00	0.00	0.50	0.0
1.00	0.14	0.50	0.2
1.00	0.35	0.50	0.5
1.00	0.57	0.50	0.8
1.00	0.00	0.75	0.0
1.00	0.17	0.75	0.2
1.00	0.43	0.75	0.5
1.00	0.69	0.75	0.8
1.00	0.00	1.00	0.0
1.00	0.20	1.00	0.2
1.00	0.50	1.00	0.5
1.00	0.80	1.00	0.8

Note: The Cov(θ,γ) was rounded up to the second decimal point.

For each condition, 100 replications were made and results were summarized with regard to the recovery of the target dimension variance σ^2_θ, the testlet effect $\sigma^2_{\gamma_d}$, and the item parameters. In addition, rejection rates from the likelihood ratio test between the regular and extended testlet models were investigated in order to determine which model was more statistically fit.

A standard method was used for item response generation. The steps for data generation were: (1) generate θ and γ values from Equation 8 following the simulation conditions shown in Table 1; (2) calculate the IRF values using the results of step (1) and the item parameter values; and (3) compare each value from step (2) with a random number drawn from a standard uniform distribution, assigning 1 when the random number is less than or equal to the IRF value and 0 otherwise.

Summary Statistics

For the recovery of the Σ matrix (target dimension variance, testlet effect, and covariance) and the item parameters, bias, root mean squared error (RMSE), and mean absolute difference (MAD) were used. Definitions of these indices follow below:

$$\text{Bias} = \frac{\sum_{r=1}^{R} \hat{\xi}_r}{R} - \xi, \tag{9}$$

where ξ is the parameter of interest, $\hat{\xi}$ is the estimate of the parameter, and R is the number of replications, and:

$$\text{RMSE} = \sqrt{\frac{\sum_{r=1}^{R} (\hat{\xi}_r - \xi)^2}{R}}, \tag{10}$$

$$\text{MAD} = \frac{\sum_{r=1}^{R} \left| \hat{\xi}_r - \xi \right|}{R}. \tag{11}$$

For the item parameters, descriptive statistics such as average across individual items were used to summarize the results.

Lastly, rejection rates from the likelihood ratio test comparing the regular and the extended Rasch testlet models were investigated for each of the 16 conditions by counting the number of statistically significant results at the $\alpha = 0.05$ level.

Results

Bias

The bias for the elements in the $\hat{\Sigma}$ matrix was shown in Table 2. One obvious result was the positive bias for the target dimension variance σ^2_θ in the regular testlet model. The maximum bias for σ^2_θ in the regular testlet model was 0.46. When θ and γ were independent ($Cov(\theta, \gamma) = 0$), which is the assumption of the regular testlet model, the bias for σ^2_θ in the regular testlet model was less than 0.02. However, when $Cov(\theta, \gamma)$ was non-zero, the bias for σ^2_θ in the regular testlet model increased with the level of true dimension correlation $\rho_{(\theta,\gamma)}$. The bias for σ^2_θ in the regular testlet model also increased as the true testlet effect $\sigma^2_{\gamma_d}$ increased. The testlet effect $\sigma^2_{\gamma_d}$ in the regular testlet model tended to increase as the true dimension correlation increased, but its bias was generally much smaller than the bias for σ^2_θ.

In the extended testlet model, the target dimension variance σ^2_θ and the testlet effect $\sigma^2_{\gamma_d}$ showed positive bias while the covariance estimate showed negative bias. The bias for σ^2_θ and $\sigma^2_{\gamma_d}$ ranged from zero to 0.09. The bias for covariance ranged from –0.01 to –0.04.

There were no particular irregularities in the bias values for the item parameters. Table 3 summarizes these item parameter biases. The average and median values of the bias were both close to zero, but the standard deviation and the range of the bias increased as the dimensional correla-

Table 2

Bias for Variances and Covariance of Random Parameters

True Σ for each condition				Bias					
				Regular Testlet Model			Extended Testlet Model		
σ^2_θ	$Cov(\theta,\gamma)$	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	$\hat{\sigma}^2_\theta$	$Cov(\theta,\gamma)$	$\hat{\sigma}^2_{\gamma_d}$	$\hat{\sigma}^2_\theta$	$Cov(\theta,\gamma)$	$\hat{\sigma}^2_{\gamma_d}$
1.00	0.00	0.25	0.00	0.00	NA	0.03	0.02	–0.02	0.03
1.00	0.10	0.25	0.20	0.07	NA	0.05	0.01	–0.02	0.03
1.00	0.25	0.25	0.50	0.20	NA	0.09	0.03	–0.02	0.05
1.00	0.40	0.25	0.80	0.32	NA	0.10	0.05	–0.04	0.08
1.00	0.00	0.50	0.00	0.00	NA	0.01	0.00	–0.01	0.02
1.00	0.14	0.50	0.20	0.09	NA	0.06	0.01	–0.01	0.02
1.00	0.35	0.50	0.50	0.24	NA	0.10	0.02	–0.02	0.03
1.00	0.57	0.50	0.80	0.40	NA	0.09	0.04	–0.04	0.07
1.00	0.00	0.75	0.00	0.00	NA	0.00	0.01	–0.01	0.01
1.00	0.17	0.75	0.20	0.11	NA	0.07	0.01	0.00	0.02
1.00	0.43	0.75	0.50	0.26	NA	0.09	0.01	–0.02	0.02
1.00	0.69	0.75	0.80	0.43	NA	0.10	0.03	–0.03	0.09
1.00	0.00	1.00	0.00	0.02	NA	0.01	0.02	–0.01	0.01
1.00	0.20	1.00	0.20	0.11	NA	0.05	0.00	–0.01	0.00
1.00	0.50	1.00	0.50	0.29	NA	0.06	0.01	–0.01	0.00
1.00	0.80	1.00	0.80	0.46	NA	0.03	0.04	–0.04	0.04

Note: Bias was rounded up to the second decimal point. NA = not applicable because it is fixed as zero.

tion $\rho_{(\theta,\gamma)}$ increased in the regular testlet model, indicating that higher $\rho_{(\theta,\gamma)}$ may create somewhat more instability than usual in the item parameter estimates in the regular testlet model.

Root Mean Squared Error (RMSE)

Table 4 provides the RMSEs of the variance and the covariance. Figure 1 summarizes the behavior of the RMSEs across different values of target dimension variance σ^2_θ and testlet effect $\sigma^2_{\gamma_d}$. In Figure 1, regular testlet model (RT) results were represented as circles connected by solid lines while extended testlet model (ET) results were represented as triangles connected by dotted lines. The RMSE values for the regular testlet model showed an obvious pattern of increase for the target dimension variance σ^2_θ and the testlet effect $\sigma^2_{\gamma_d}$ as the true dimension correlation increased. The larger true testlet effect also

Table 3

Summary of Item Bias

True Σ for each condition				Item Parameter Bias Summary					
						Standard			
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	Median	Average	Deviation	Min	Max	Range
				The Regular Testlet Model					
1.00	0.00	0.25	0.00	0.00	0.01	0.01	–0.02	0.03	0.05
1.00	0.10	0.25	0.20	0.01	0.00	0.01	–0.02	0.03	0.05
1.00	0.25	0.25	0.50	0.01	0.00	0.02	–0.04	0.03	0.07
1.00	0.40	0.25	0.80	0.02	0.01	0.03	–0.05	0.07	0.12
1.00	0.00	0.50	0.00	0.00	0.00	0.01	–0.02	0.02	0.03
1.00	0.14	0.50	0.20	0.01	0.01	0.02	–0.02	0.04	0.06
1.00	0.35	0.50	0.50	0.00	–0.01	0.02	–0.06	0.04	0.10
1.00	0.57	0.50	0.80	0.01	0.00	0.04	–0.10	0.07	0.17
1.00	0.00	0.75	0.00	0.01	0.01	0.01	–0.01	0.02	0.03
1.00	0.17	0.75	0.20	0.00	0.01	0.01	–0.02	0.03	0.05
1.00	0.43	0.75	0.50	0.01	0.01	0.03	–0.06	0.06	0.12
1.00	0.69	0.75	0.80	0.01	0.01	0.05	–0.08	0.10	0.18
1.00	0.00	1.00	0.00	0.00	0.00	0.01	–0.01	0.02	0.03
1.00	0.20	1.00	0.20	0.00	0.00	0.01	–0.03	0.03	0.05
1.00	0.50	1.00	0.50	0.01	0.00	0.03	–0.06	0.05	0.12
1.00	0.80	1.00	0.80	0.01	0.00	0.05	–0.10	0.09	0.19
				The Extended Testlet Model					
1.00	0.00	0.25	0.00	0.01	0.01	0.01	–0.02	0.03	0.05
1.00	0.10	0.25	0.20	0.00	0.00	0.01	–0.01	0.02	0.04
1.00	0.25	0.25	0.50	0.00	0.00	0.01	–0.02	0.01	0.03
1.00	0.40	0.25	0.80	0.01	0.01	0.01	–0.01	0.02	0.03
1.00	0.00	0.50	0.00	0.00	0.00	0.01	–0.02	0.02	0.03
1.00	0.14	0.50	0.20	0.01	0.01	0.01	–0.01	0.02	0.03
1.00	0.35	0.50	0.50	–0.01	–0.01	0.01	–0.02	0.01	0.03
1.00	0.57	0.50	0.80	0.00	0.00	0.01	–0.02	0.01	0.03
1.00	0.00	0.75	0.00	0.01	0.01	0.01	–0.01	0.02	0.02
1.00	0.17	0.75	0.20	0.00	0.00	0.01	–0.02	0.02	0.04
1.00	0.43	0.75	0.50	0.00	0.00	0.01	–0.01	0.02	0.03
1.00	0.69	0.75	0.80	0.00	0.00	0.01	–0.03	0.02	0.04
1.00	0.00	1.00	0.00	0.00	0.00	0.01	–0.01	0.02	0.03
1.00	0.20	1.00	0.20	0.00	0.00	0.01	–0.02	0.01	0.03
1.00	0.50	1.00	0.50	–0.01	0.00	0.01	–0.02	0.01	0.03
1.00	0.80	1.00	0.80	0.00	–0.01	0.01	–0.02	0.00	0.02

increased the RMSE values. The RMSE for σ^2_θ ranged from 0.004 to 0.227 in the regular testlet model. The RMSE for $\sigma^2_{\gamma_d}$ in the regular testlet model was generally much smaller (minimum RSME = 0.002; maximum = 0.021) than that of σ^2_θ. In the regular testlet model, the maximum RMSE for σ^2_θ was about 11 times larger than the maximum RMSE for $\sigma^2_{\gamma_d}$.

The testlet effect $\sigma^2_{\gamma_d}$ in the extended testlet model also showed increment in its RMSE as the true dimension correlation and the size of the true testlet effect increased. In general, however, the RMSE for $\sigma^2_{\gamma_d}$ were smaller in the extended testlet model (the minimum = 0.002; the maximum = 0.018) than in the regular testlet model. The target variance σ^2_θ and covariance in the extended testlet model did not show any peculiar patterns.

RMSEs for the item parameters were summarized using the average of all item RMSEs for each condition. Table 5 presents these average RMSE values.

The average RMSE values for the regular testlet model tended to increase as the true dimensional correlation increased. The extended testlet model showed equal or slightly better performance with regard to average RMSE (maximum difference of 0.002 compared to the regular testlet model).

Mean Absolute Difference (MAD)

The RMSE statistic is an overall measure of accuracy and stability for parameter recovery. However, RMSE is less easy to interpret than the MAD statistic, which simply represents the average absolute difference between the estimate and the true value. MAD values for the variances and the covariance are shown in Table 6. The overall pattern of the MAD results was very similar to that of the RMSE results. Figure 2 presents a summary of the MAD values.

The MAD values for the target variance σ^2_θ in the regular testlet model increased as the true dimensional correlation increased (the maximum MAD was 0.49 and the minimum MAD was 0.05). MAD also increased with the increment of the size of the true testlet effect. The testlet effect $\sigma^2_{\gamma_d}$ in the regular testlet model showed a similar pattern of increasing MAD as true dimension cor-

Table 4

Root Mean Squared Error (RMSE) for the variances and the covariance

True Σ for each condition				RMSE					
				Regular Testlet Model			Extended Testlet Model		
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	$\hat{\sigma}^2_\theta$	Cov(θ,γ)	$\hat{\sigma}^2_{\gamma_d}$	$\hat{\sigma}^2_\theta$	Cov(θ,γ)	$\hat{\sigma}^2_{\gamma_d}$
1.00	0.00	0.25	0.00	0.004	NA	0.002	0.007	0.004	0.003
1.00	0.10	0.25	0.20	0.010	NA	0.005	0.007	0.002	0.003
1.00	0.25	0.25	0.50	0.045	NA	0.012	0.005	0.003	0.006
1.00	0.40	0.25	0.80	0.113	NA	0.014	0.009	0.004	0.009
1.00	0.00	0.50	0.00	0.004	NA	0.005	0.006	0.003	0.005
1.00	0.14	0.50	0.20	0.014	NA	0.010	0.006	0.004	0.007
1.00	0.35	0.50	0.50	0.065	NA	0.016	0.007	0.004	0.007
1.00	0.57	0.50	0.80	0.164	NA	0.015	0.006	0.005	0.010
1.00	0.00	0.75	0.00	0.005	NA	0.008	0.006	0.004	0.008
1.00	0.17	0.75	0.20	0.018	NA	0.013	0.006	0.004	0.009
1.00	0.43	0.75	0.50	0.076	NA	0.018	0.005	0.005	0.011
1.00	0.69	0.75	0.80	0.191	NA	0.021	0.006	0.004	0.018
1.00	0.00	1.00	0.00	0.005	NA	0.010	0.007	0.004	0.011
1.00	0.20	1.00	0.20	0.018	NA	0.014	0.006	0.003	0.010
1.00	0.50	1.00	0.50	0.088	NA	0.017	0.005	0.005	0.013
1.00	0.80	1.00	0.80	0.227	NA	0.014	0.008	0.005	0.016

Note: The RMSE was rounded up to the third decimal point to show differences.

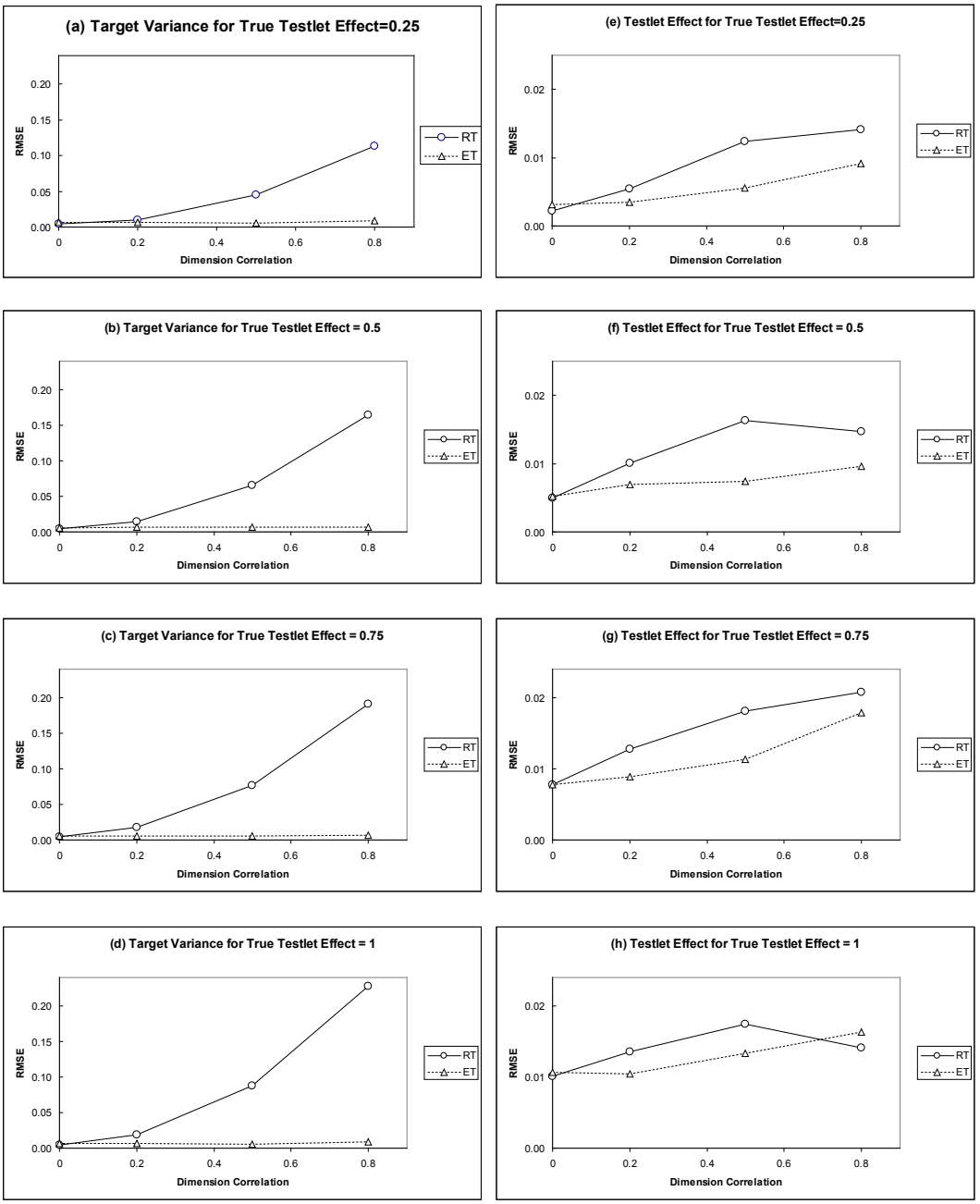

Figure 1. RMSE for the Target Variance and the Testlet Effect
Note: RT represents the regular testlet model and ET represents the extended testlet model.

relation increased. In the independent distribution condition in which covariance is zero, the regular testlet model performed better than the extended testlet model only by a slight margin (MAD values were lower for the regular testlet model by less than about 0.01). Under all conditions with non-zero covariances, the extended testlet model showed better performance: for estimates of the target variance σ^2_θ, the maximum difference in MAD between the regular and extended models

Table 5

Average RMSE for Item Parameters

True Σ for each condition				Average RMSE	
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	Regular Testlet Model	Extended Testlet Model
1.00	0.00	0.25	0.00	0.007	0.007
1.00	0.10	0.25	0.20	0.007	0.007
1.00	0.25	0.25	0.50	0.008	0.008
1.00	0.40	0.25	0.80	0.009	0.008
1.00	0.00	0.50	0.00	0.007	0.007
1.00	0.14	0.50	0.20	0.008	0.008
1.00	0.35	0.50	0.50	0.007	0.007
1.00	0.57	0.50	0.80	0.009	0.007
1.00	0.00	0.75	0.00	0.007	0.007
1.00	0.17	0.75	0.20	0.008	0.008
1.00	0.43	0.75	0.50	0.008	0.008
1.00	0.69	0.75	0.80	0.010	0.008
1.00	0.00	1.00	0.00	0.007	0.008
1.00	0.20	1.00	0.20	0.008	0.008
1.00	0.50	1.00	0.50	0.009	0.008
1.00	0.80	1.00	0.80	0.011	0.009

Note: The RMSE was rounded up to the third decimal point to show differences.

Table 6

Mean Absolute Difference (MAD) for the variances and the covariance

True Σ for each condition				MAD					
				Regular Testlet Model			Extended Testlet Model		
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	$\hat{\sigma}^2_\theta$	Cov(θ,γ)	$\hat{\sigma}^2_{\gamma_d}$	$\hat{\sigma}^2_\theta$	Cov(θ,γ)	$\hat{\sigma}^2_{\gamma_d}$
1.00	0.00	0.25	0.00	0.05	NA	0.04	0.06	0.05	0.04
1.00	0.10	0.25	0.20	0.09	NA	0.06	0.06	0.04	0.05
1.00	0.25	0.25	0.50	0.20	NA	0.10	0.06	0.04	0.06
1.00	0.40	0.25	0.80	0.32	NA	0.10	0.07	0.05	0.08
1.00	0.00	0.50	0.00	0.05	NA	0.06	0.06	0.04	0.06
1.00	0.14	0.50	0.20	0.10	NA	0.08	0.06	0.05	0.07
1.00	0.35	0.50	0.50	0.24	NA	0.11	0.06	0.05	0.07
1.00	0.57	0.50	0.80	0.40	NA	0.10	0.06	0.05	0.08
1.00	0.00	0.75	0.00	0.06	NA	0.07	0.06	0.05	0.07
1.00	0.17	0.75	0.20	0.11	NA	0.09	0.06	0.05	0.08
1.00	0.43	0.75	0.50	0.26	NA	0.10	0.06	0.06	0.08
1.00	0.69	0.75	0.80	0.43	NA	0.12	0.07	0.05	0.11
1.00	0.00	1.00	0.00	0.06	NA	0.08	0.07	0.05	0.08
1.00	0.20	1.00	0.20	0.12	NA	0.09	0.06	0.05	0.08
1.00	0.50	1.00	0.50	0.29	NA	0.10	0.06	0.06	0.10
1.00	0.80	1.00	0.80	0.46	NA	0.09	0.07	0.06	0.10

Note: The MAD values were rounded up to the second decimal point.

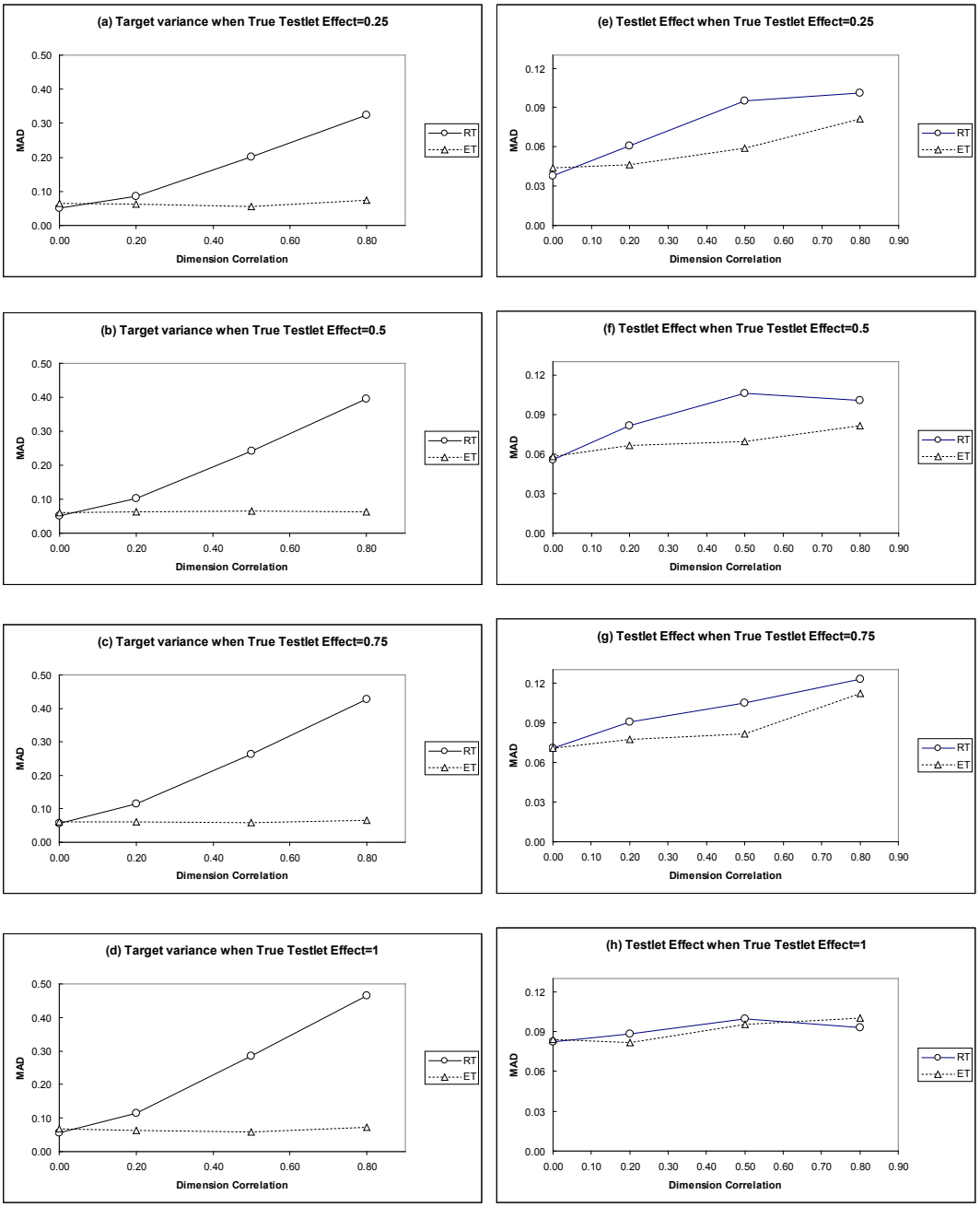

Figure 2. MAD for the Target Variance and the Testlet Effect
Note: RT represents the regular testlet model and ET represents the extended testlet model.

was 0.391; and for estimates of the testlet effect $\sigma^2_{\gamma_d}$ the maximum difference was close to 0.04. MAD values for the covariance in the extended testlet model ranged from about 0.04 through 0.06. The testlet effect $\sigma^2_{\gamma_d}$ for the regular and extended models was estimated very closely to each other when the true testlet effect was 1 and the difference in MAD between the two models was less than 0.01.

MAD for the item parameters was summarized by averaging across all items under each set of conditions. Table 7 shows these average MAD values.

As was the case for RMSE, MAD values were the same for both the regular and the extended testlet models or were slightly lower for the extended model (differences in MAD between the models ranged from virtually zero to 0.009). As the true dimensional correlation $\rho_{(\theta,\gamma)}$ increased, the MAD for item parameters increased as well. This effect was more salient for the regular testlet model than for the extended testlet model.

Model Comparison by Likelihood Ratio Test

The regular Rasch testlet model is a special case of the extended Rasch testlet model, since the latter can be reduced to the former by fixing the covariance(s) between the random parameters at zero. Due to this nested structure, statistical comparisons between the models by means of a likelihood ratio were possible. The null hypothesis was the regular Rasch testlet model and the alternative was the extended Rasch testlet model (this formulation is equivalent to testing the null hypothesis of zero covariance).

The rejection rates under $\alpha = 0.05$ are presented for all conditions in Table 8. For the conditions of the covariance between θ and γ equal to zero, corresponding to the assumption of the regular testlet model, the null hypothesis of the regular testlet model was rejected 3% to 7% of the time, close to what would be expected to occur by chance under 0.05 level. As the true dimension correlation increased and as the size of the true testlet effect became large, the rejection rate increased, i.e., the extended testlet model was preferred to the regular testlet model more often. When the true dimension correlation was greater than or equal to 0.5, the extended testlet model was preferred to the regular model in virtually all instances, regardless of the size of the true testlet effect.

Table 7

Average MAD for item parameters

True Σ for each condition				Average MAD	
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	Regular Testlet Model	Extended Testlet Model
1.00	0.00	0.25	0.00	0.066	0.066
1.00	0.10	0.25	0.20	0.066	0.066
1.00	0.25	0.25	0.50	0.072	0.070
1.00	0.40	0.25	0.80	0.073	0.069
1.00	0.00	0.50	0.00	0.066	0.066
1.00	0.14	0.50	0.20	0.071	0.071
1.00	0.35	0.50	0.50	0.068	0.066
1.00	0.57	0.50	0.80	0.075	0.067
1.00	0.00	0.75	0.00	0.067	0.067
1.00	0.17	0.75	0.20	0.070	0.069
1.00	0.43	0.75	0.50	0.073	0.070
1.00	0.69	0.75	0.80	0.080	0.072
1.00	0.00	1.00	0.00	0.069	0.069
1.00	0.20	1.00	0.20	0.072	0.071
1.00	0.50	1.00	0.50	0.076	0.074
1.00	0.80	1.00	0.80	0.081	0.074

Note: The average MAD values were rounded up to the third decimal point to show differences.

Table 8

Model Comparison Results from Likelihood Ratio Test

True Σ for each condition				Model Comparison by LR test	
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	Non-Rejection Rate (%)	Rejection Rate (%)
1.00	0.00	0.25	0.00	94	6
1.00	0.10	0.25	0.20	63	37
1.00	0.25	0.25	0.50	2	98
1.00	0.40	0.25	0.80	0	100
1.00	0.00	0.50	0.00	94	6
1.00	0.14	0.50	0.20	38	62
1.00	0.35	0.50	0.50	1	99
1.00	0.57	0.50	0.80	0	100
1.00	0.00	0.75	0.00	93	7
1.00	0.17	0.75	0.20	19	81
1.00	0.43	0.75	0.50	0	100
1.00	0.69	0.75	0.80	0	100
1.00	0.00	1.00	0.00	97	3
1.00	0.20	1.00	0.20	18	82
1.00	0.50	1.00	0.50	0	100
1.00	0.80	1.00	0.80	0	100

Summary and Discussion

In some parametric IRT testlet models that have been proposed previously, regardless of whether the approach was Bayesian (e.g., Wang, et al., 2002) or non-Bayesian (e.g., Wang and Wilson, 2005), the person ability parameter and the testlet effect parameter were treated as random parameters and were independently and normally distributed. This article investigated the effect of violating the assumption of independence between these random parameters (the target dimension θ and the testlet effect γ) and tested the performance of an extended Rasch testlet model in which the variance-covariance matrix elements for the random parameters were freely estimated. Major findings for the regular Rasch testlet model were as follows. When the assumption of independence was violated by introducing a positive correlation between the two random parameters, (1) the target ability dimension variance σ^2_θ was strongly affected (the maximum difference from the true value across all conditions was close to 0.5); (2) the testlet effect $\sigma^2_{\gamma_d}$ and item parameters were much less influenced than the target dimension variance (the largest absolute differences across all conditions were less than 0.12 for $\sigma^2_{\gamma_d}$ and the item parameters); and (3) the increase in bias, RMSE, and MAD values for the target dimension variance was a function of the true dimensional correlation $\rho_{(\theta,\gamma)}$ (or covariance) and the size of the true testlet effect. The performance of the extended Rasch testlet model was superior or approximately equal to that of the regular Rasch testlet model in terms of bias, RMSE, and MAD in all conditions. To facilitate interpretation of the comparisons between the performances of the regular and extended testlet models, percent relative error (PRE) was calculated and presented in Table 9. PRE was defined as 100 $\times$ (average estimate for the regular testlet model – average estimate for the extended testlet model)/(average estimate for the extended testlet model).

PRE is a signed measure of percentage bias relative to the extended testlet model. For example, a PRE of 16.8 means that the estimate of the parameter of interest for the regular testlet model was higher by 16.8% than the estimate for the extended testlet model. A value of –1.4 implies that the estimate for the regular testlet model was lower by 1.4% than the estimate for the extended testlet model. Table 9 shows that overall, the regular test model produced higher estimates for the target dimension variance and the testlet effect than did the extended testlet model, arriving at estimates up to 41% higher than those from the

Table 9

Percent Relative Error for the Regular Testlet Model

True Σ for each condition				PRE	
σ^2_θ	Cov(θ,γ)	$\sigma^2_{\gamma_d}$	$\rho_{(\theta,\gamma)}$	RT σ^2_θ	RT $\sigma^2_{\gamma_d}$
1.00	0.00	0.25	0.00	–1.4	–2.5
1.00	0.10	0.25	0.20	6.3	7.4
1.00	0.25	0.25	0.50	16.8	12.9
1.00	0.40	0.25	0.80	26.1	5.8
1.00	0.00	0.50	0.00	–0.6	–0.7
1.00	0.14	0.50	0.20	8.7	7.4
1.00	0.35	0.50	0.50	22.2	11.9
1.00	0.57	0.50	0.80	34.3	4.2
1.00	0.00	0.75	0.00	–0.7	–0.7
1.00	0.17	0.75	0.20	10.7	6.3
1.00	0.43	0.75	0.50	24.9	9.0
1.00	0.69	0.75	0.80	38.5	1.6
1.00	0.00	1.00	0.00	–0.6	–0.5
1.00	0.20	1.00	0.20	10.7	5.0
1.00	0.50	1.00	0.50	26.8	6.3
1.00	0.80	1.00	0.80	40.7	–1.1

Note: PRE is defined as 100x (RT average estimate – ET average estimate)/(ET average estimate). It shows bias direction and the amount of bias as a percentage relative to the extended testlet model.

extended model. The pattern of overestimation by the regular testlet model relative to the extended testlet model was again evident for the target dimension variance.

The amount of bias calculated for the target dimension variance suggests a caveat for the interpretation of the estimated testlet effect in the regular Rasch testlet model. If the size of testlet effect is expressed as a percentage of the target dimension variance:

$$100x\frac{\sigma^2_{\gamma d}}{\sigma^2_\theta},$$

and if $\sigma^2_{\gamma_d}$ is estimated to be less positively biased than σ^2_θ as this study simulation showed, then the percentage ratio will tend to be underestimated and the size of the testlet effect can be misjudged. In reality, when the regular testlet model is applied, only σ^2_θ and $\sigma^2_{\gamma_d}$ are estimated and there is no easy way to determine the degree of correlation between the dimensions. The extended Rasch testlet model has a potential to correct this problem. When using the extended testlet model, dimension covariances are estimated and potential biases for target variance(s), testlet effects, and item parameters that may occur in the regular testlet model are expected to be reduced.

The regular and extended Rasch testlet models can be understood as applications of the multidimensional Rasch model. Hence, this study can be seen as a robustness analysis of a particular multidimensional Rasch model from a multidimensional IRT modeling point of view. However, before the regular and extended Rasch testlet models can be fully used in practical applications, more simulations with a variety of conditions and real data application research appear to be required.

There are many limitations in this study. To name a few, first of all, for clarity, the research here assumed a simple structure for the random parameters (i.e., bivariate normal distribution between θ and γ). Currently there is no extensive comparative information available as to how the regular and extended testlet models behave when applied to a non-orthogonal multivariate distribution structure (e.g., more testlets and non-zero correlations among the target dimen-

sion and the testlet effects with varying number of items per testlet, which is much more complex and yet closer to reality). And this study did not include a real-data application, which may shed further light on how correlations between the target dimension and the testlet effect(s) should be interpreted and to what extent they can be expected to occur in real-data applications. Further studies involving both real data and a wide range of simulations with diverse testlet conditions will test and verify the utilities of the proposed extended testlet model in this study.

References

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

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Du, Z. (1998). *Modeling conditional item dependencies with a three-parameter logistic testlet model*. Unpublished doctoral dissertation, Columbia University, New York City.

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: MESA Press.

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *14*, 59-71.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, *63*, 349-359.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *17*, 1-100.

Wainer, H., and Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-202.

Waier, H., Bradlow, E. T., and Du, Z. (2000). Testlet response theory: An analog for the 3PL model using in testlet-based adaptive testing. In W. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). London: Kluwer.

Wainer, H., and Wang, Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*, 22-29.

Wainer, H., and Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203-220.

Wang, X., Bradlow, E. T., and Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*, 109-128.

Wang, W.-C., and Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement*, *29*, 126-149.

Wilson, M. R., and Adams, R. J. (1995). Rasch models for item bundles. *Psychometrika*, *60,* 181-198.

Wu. M. L., Adams, R. J., and Wilson, M. R. (1998). ConQuest: Generalized item response modeling software [Computer program]. Camberwell, VIC, Australia: Australian Council for Educational Research.