Applications of the Testlet Response Model to Performance Simulation Tasks

Jonathan Rubright, MS

University of Delaware

Michael S. Finger, PhD

American Institute of Certified Public Accountants

Applications of the Testlet Response Model to Performance Simulation Tasks

Item Response Theory (IRT) models define, in probabilistic terms, what is expected to happen when an examinee with a given ability encounters an item with a given set of characteristics. One assumption of IRT models is that of local item dependence (LID) (Hambleton, Swaminathan, & Rogers, 1991); since estimation requires only examinee ability and item characteristics to model a response, IRT models assume that responses between any items will be independent after controlling for the overall ability of interest. This independence can be thought of in numerous ways. In factor analytic language, after removing the general factor of interest, the items are not expected to have any remaining residual correlations. Or, the conditional probability of an examinee's joint distribution of scores on any two items should equal the product of probabilities for those items, such that θ provides all the information necessary about an examinee so each item can be treated independently (W Yen, 1993):

(1) $$P(X_1 = x_1, X_2 = x_2 \mid \theta) = P(X_1 = x_1 \mid \theta)P(X_2 = x_2 \mid \theta)$$

However, this assumption may not always hold true. Not infrequently, items are grouped together in response to a common stimulus. For example, a series of reading comprehension questions may be asked about a passage of literature. Or, an examinee may be asked to interpret a graph or chart. Statistical considerations aside, these situations appear, on the face, to violate local independence. Questions referring to a common stimulus might be expected to share at least some dependency. An examinee may have previously seen the stimulus or have particular expertise in that area, and therefore may get more items correct than their general ability would otherwise predict. Conversely, misinterpreting a common stimulus may make an examinee more likely to

get the associated items incorrect. Additional plausible scenarios for violating local item independence have been enumerated elsewhere (W Yen, 1993). Of course, statistical estimates of dependency, such as Q_3, can be calculated to describe the amount of any dependency present (Y. Lee, 2004).

Formally, these groups of items have been called testlets, “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow" (p. 190) (Wainer & Kiely, 1987). Having this arrangement of items as a part of a test seems appropriate for a number of reasons. First, it can increase the face validity of a test so that it more closely resembles real-world tasks. Thus, testlets may “reduce concerns about the atomistic nature of single independent small items” (p. 246) (Wainer, Bradlow, & Du, 2000). Secondly, in situations where response to a passage or chart is desirable, presenting a stimulus and asking only a single question would considerably increase the length of a test, as a number of stimuli and related items would be needed to generate enough responses from an examinee. This is problematic since much research into the design of computerized tests has explicitly attempted to shorten test length.

On tests where dependency may be a concern, at least two questions present themselves. Since IRT models make the assumption of local item independence, what is the effect of fitting these models to data where this assumption does not hold? Secondly, and perhaps more importantly, how can tests with dependence be most appropriately modeled?

Studies have evidenced a number of concerns when fitting unidimensional IRT models to data from tests with known dependencies. Some have provided theoretical

reasoning why testlets overestimate reliability or information (Sireci, Wainer, & Thissen, 1991; W Yen, 1993). Others have shown overestimates of reliability using unidimensional models when dependencies are expected on tests such as the LSAT (Wainer, 1995), the SAT-V (Sireci et al., 1991), and TOEFL (Wainer & Lukhele, 1997). TOEFL data also suggests that while difficulty parameters can be fairly well estimated using unidimensional models, discrimination and guessing parameters can be seriously biased (Wainer & Wang, 2000).

Simulation studies, where local item dependences can be controlled and item parameters known, also provide data on the impact of dependencies on parameter estimates. Studies of this type reaffirm results from real datasets, showing that the use of unidimensional models on data generated with dependencies have underestimates of discrimination parameters, higher Root Mean Square Errors (RMSEs) for ability estimates, underestimates of standard errors of measurement, bias in conditional standard errors of measurement, and overestimates of test information and reliability (Bradlow, Wainer, & Wang, 1999; DeMars, 2006; G. Lee, 2000; Wainer et al., 2000). As expected, bias in estimates worsens as the degree of dependency increases (Ackerman, 1987; DeMars, 2006; G. Lee, 2000). Thus, the further a test departs from unidimensionality, the more flawed estimates from unidimensional models become.

In light of evidence that ignoring local dependence skews estimates, a few ways of dealing with dependence have been considered. The first is to simply ignore the dependence and apply unidimensional IRT models. Yet, particularly when dependence is high, bias will result (as seen above). Secondly, a test could present only one item after each stimulus in order to avoid dependence altogether. As mentioned above, this will

lengthen the test and increase associated costs in time and resources. Third, one could use an alternative model. As recently as 1996, many (Sireci et al., 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1995; Wainer & Thissen, 1996) had advocated use of polytomous models to account for testlet dependency.

For example, if a reading passage was followed by 5 items, one could score each item and add up the total on a polytomous scale, with a score ranging from 0 to 5. Thus, each testlet effectively becomes a single, polytomously scored item. Subsequently, these scores could be used to fit a polytomous IRT model (Samejima, 1969) and estimated using commercially available software (Thissen, 1993). Simulation studies suggest that this technique performs better than a dichotomous model when dependencies are high, having lower RMSEs between parameter estimates and improving equating (G. Lee, Kolen, Frisbie, & Ankenmann, 2001; Stark, Chernyshenko, & Orleans., 2002).

Still, this polytomous work-around has two main shortcomings. As it is based on a simple, additive number correct, the polytomous IRT model cannot take advantage of the response pattern within the testlet and necessarily utilizes less information. Secondly, it cannot be used in adaptive settings where individuals may have different follow-up questions within a testlet (Wainer, Bradlow, & Wang, 2007).

More recently, newer models have been developed to explicitly model and account for the additional dependence between items within a testlet. These models, called Testlet Response Theory (TRT) models, are described in a book that is a summary of relevant publications on the topic (Wainer et al., 2007). The first description of the approach generated a formula that was an extension of the two parameter IRT model, adding a random effect to account for dependency. As anticipated by the authors, the

newer model better estimated item parameters using simulated data and provided more reasonable estimates for reliability using real data (Bradlow et al., 1999).

Soon thereafter, the same team developed the three parameter testlet generalization (Wainer et al., 2000) of Birnbaum's three parameter model (Birnbaum, 1968). The traditional three parameter model is calculated as:

$$(2) \qquad p(y_{ij}=1) = c_j + (1-c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1+\exp[a_j(\theta_i - b_j)]}$$

The left side is the probability of person *i* answering item *j* correctly, and on the right, θ is the ability level of person *i*, a_j is the discrimination of item *j*, b_j is the difficulty of item *j*, and c_j is the guessing parameter of item *j*. The extension by Wainer and colleagues is written as:

$$(3) \qquad p(y_{ij}=1) = c_j + (1-c_j)\frac{\exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]}{1+\exp[a_j(\theta_i - b_j - \gamma_{ig(j)})]}$$

The new TRT adds "a random effect to the logit . . . that is an interaction of person *i* with testlet *g(j)*, the testlet that contains item *j*" (p. 205) (Wainer & Wang, 2000).

In instances where γ is zero, the model simplifies to the three parameter model, clearly showing how the three parameter model is nested within the TRT model. Similarly, when the guessing parameter c_j is zero, the three parameter model simplifies to the two parameter model. Thus, both the 2PL testlet and 3PL testlet models are nested among the more traditional, unidimensional approaches. These newer models can be estimated within a fully Bayesian framework utilizing the SCORIGHT software written by the authors of the models (Wang, Bradlow, & Wainer, 2004). These TRT models are functionally related to bi-factor and higher-order, multidimensional models (de la Torre & Song, 2009; Li, Bolt, & Fu, 2006; Yung, Thissen, & McLeod, 1999). However,

notably, the TRT models have a single a parameter to characterize the relationship between the theta and the testlet gamma.

In TRT, the item is still the unit of measure as in traditional IRT modeling approaches. However, a parameter is now included to model the amount of local dependence within each testlet, and a number of authors have provided evidence that the TRT model works well in comparison to unidimensional models when dependence is present (Bradlow et al., 1999; DeMars, 2006; Du, 1998; Li & Cohen, 2003; Wainer et al., 2000; Wang, Bradlow, & Wainer, 2002). In comparison to BILOG estimates, the TRT model seems to produce better estimates, particularly in the discrimination parameters. Wainer explains, "that when there is unmodeled local dependence the model looks upon the resulting nonfit as noise and so the regression of the item on the underlying trait is . . . more gradual" (p. 265) (Wainer et al., 2000). With LD accounted for, the a parameter estimates become steeper. [Newer evidence suggests that a parameter comparisons between models are best performed after marginalizing the TRT model estimates (Ip, 2010).] In regards to test information functions, when dependency is present, is has been shown that TRT test information functions (TIFs) are more platykurtic than those generated from unidimensional models (Ip, 2010).

Interestingly, TRT models can be used in the presence or absence of dependency: when present, the models seem to provide more faithful estimates of parameters, yet when dependence is missing (when $\gamma=0$), the models produce results similar to the unidimensional models. This property makes the TRT models candidates for tests that have mixtures of testlets and independent items.

More common applications of TRT apply the model to multiple-choice (MC) items, in which sets of MC items each address some common reading passage or mathematics problem. Much less common is applying TRT models to items from performance tasks that are designed to simulate real-world conditions and assess skill levels, as opposed to only content knowledge. In simulations found on some licensure and certification examinations, an examinee is presented with a task designed to emulate a relevant and realistic, work-based scenario. The tasks are comprised of discrete actions, steps, or entries that are each treated as an item within the task and scored dichotomously (correct/incorrect). While the series of items within each task can be modeled using standard unidimensional models, these models would not account for any LID that may be present due to the common task stimulus or scenario.

In the present study, empirical simulation task data is taken from three sections of a large-scale, multi-section licensure examination in order to explore the utility of the TRT model both in terms of post-administration purposes of item and person parameter estimation. The TRT extension of the 2PL (i.e., a 2PL with an added testlet person parameter) is compared to the standard 2PL on the empirical datasets to assess the extent of overall testlet effects on the test and to compare the extent of LID based on each model. Additionally, examinee abilities, item parameter estimates, and TIFs are compared between the models to ascertain the extent to which the models generate differing results and whether the added model complexity of the TRT model provides additional information.

Methods.

Data

Item-level simulation task data are taken from three of the four sections from a large-scale, national licensure examination. Each of these sections has three components: a set of MC items, a single-scored constructed response item, and a group of 8 simulation tasks. The total score from each component are weighted, summed, and rescaled to obtain a pass/fail status based on an overall cutscore.

The simulation tasks from each of three test forms per each section were selected from a larger pool of a testing window period that spanned two months. Table 1 presents a summary of the number of forms, sample size, and number of items by form and section that are analyzed. Each form has 8 tasks, some of which have multiple items per task. Each form has 6 testlets, made up of 2 to 8 items. The total number of task items ranges from 30 – 43 across forms and sections.

Local Dependence

First, the severity of LID on these simulations is assessed. While many measures of LID have been proposed, the most common is perhaps Yen's Q_3 statistic (WM Yen, 1984). A number of authors have shown Q_3 to be a helpful descriptive statistic for identifying instances of LID, outperforming other measures (Chen & Thissen, 1997; W Yen, 1993). To calculate Q_3, one first estimates an ability parameter for each examinee. A deviation is calculated between each examinee's observed and predicted score on each item using this ability estimate. For each item, Q_3 is the correlation between these unique item pair residuals across the sample. For items j and j',

(4) $Q_{3jj'}=r_{(dj,dj')}$.

The obtained Q_3 value is compared to the expected value when no LID exists, defined as $-1/(n-1)$, with n being the number of test items. This study calculates the

average Q_3 values on all unique item pairs within each testlet. The theta values used to calculate Q_3 are obtained from both the unidimensional and TRT 2PL models. Unique item pair correlations within each testlet are transformed via Fisher's *z* transformations, averaged, then converted back o the correlation metric. The extent of Q_3 values is assessed, examined both within as well as across tasks. Additionally, the testlet variance [i.e., testlet effect; var(γ)] will be estimated under the TRT model. Greater values of var(γ) for a task indicate higher levels of LID.

Test and Item Parameters

To assess the impact of using the TRT model versus a standard unidimensional model, models are fit to each form individually. Due to the small values of guessing parameters present in this data, the standard 2PL model is estimated via MMLE using BILOG-MG (Mislevy & Bock, 1991) and again via MCMC utilizing the SCORIGHT software (Wang et al., 2004). The 2PL-Testlet model is estimated via SCORIGHT. All SCORIGHT models were run with 5 chains, each chain allowed 25,000 burn-in iterations, with every other draw from the next 15,000 iterations selected to form the posterior distribution of interest. Reasonable convergence was assessed by ensuring all Gelman-Rubin F-tests were well below the suggested 1.2 (Gelman & Rubin, 1993).

Since SCORIGHT utilizes the logistic metric, all a parameters are divided by 1.7 to bring them in line with the normal metric. For each form, the item parameter estimates and ability estimates of the 2PL SCORIGHT and 2PL SCORIGHT Testlet models are equated to the scale of the estimates from the 2PL BILOG-MG model using the mean/sigma approach. Item and ability estimates between pairs of models and estimation

methods are compared using Root Mean Square Difference (equation 4), Average Signed Difference (equation 5), and Pearson correlations.

(5) $$\sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_j)^2}{n}}$$

(6) $$\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_j)}{n}$$

Additionally, the TIFs are examined for each panel within each form; the function from the standard BILOG-MG 2PL model is compared to that from the 2PL-Testlet model. The information function for the TRT models are computed using the marginalized approach (Ip, 2010) as opposed to the approach suggest by the TRT authors (Wainer et al., 2000). The information functions are presented as relative efficiencies (REs), the ratios of the two information values at each point of theta, Information(TRT) / Information(2PL). RE > 1 means more test information from the TRT, and 0<RE<1 means less test information from the TRT vs. the 2PL. RE functions are plotted by section since the curves are on the same relative metric.

Results.

As revealed by the Q_3 values in Table 2, all testlets of all the panels analyzed here have higher residual correlations than would be expected if the assumption of local item independence held true. Utilizing the SCORIGHT software to implement the TRT models, Tables 3 and 4 display the Gelman-Rubin F-tests for convergence of item parameters and testlet variance estimates. The large number of iterations used here, in order to avoid issues of converge, were successful given the proximately of all convergence statistics to 1.00. To confirm convergence of models using the MCMC

parameters, additional models were run using varied posterior gap sizes. These models produced estimates similar to the third decimal place to the parameters used in this paper, giving confidence in the convergence of the SCORIGHT models. Table 5 displays Q_3 statistics using the TRT model. While many average values decrease as expected, a nontrivial number remain the same or increase.

Additional evidence for testlet effects on these tests is given by the gamma variances displayed in Table 6. In tests without such testlets, the variance would be expected to be 0. Here, the variances vary, yet are consistently above 0 for all datasets. Descriptive statistics for parameters estimated through the three models are shown in Table 7. To compare these estimates, we use Pearson correlations (Table 8), RMSD, and ASD (Table 9). Using the unidimensional 2PL estimated through BILOG-MG as a baseline, all parameters show very high, significant correlations. The only parameter correlating comparatively weakly is the a parameter for the testlet model.

Examining Table 9, it is noted that the ASD for the b and theta values are all 0, showing the successful use of the mean/sigma equating approach. The ASD and RMSD statistics for the 2PL testlet model are consistently higher than the SCORIGHT 2PL model. Figures 1 through 3 show the REs at each point of theta. Consistently over the middle range of the theta scale, the TRT REs provide less information than the unidimensional counterpart.

Conclusions.

The main question to be answered is whether performance-based task simulations carry along with them the local dependence seen in testlets made of MC items. Here, the results mirror those found in the literature when testlet effects are present, suggesting the

use of a higher-order model which can account for these violations of the LID assumption. The uniformly high Q_3 values are the first evidence for dependency on these tests. Interestingly, use of the TRT model did not, as was expected, evenly lower these values. Many remained similar or increased. This does call for a deeper examination on the change in Q_3 statistics using different models.

The large gamma variances provide further evidence of the testlet effect, along with poor correlations between BILOG-MG and the TRT a parameter estimates. While the difficulty parameter and theta estimates correlated highly, the marginalized TRT a parameters appear larger than the MMLE estimates, a result also seen in simulation studies on data generated with dependencies. These a parameters correlate poorly with those estimated using MMLE and show RMSD and ASD differences. Notably, the SCORIGHT 2PL model worked fairly well in estimating similar values to those obtained through BILOG-MG. Examining the RE figures, TRT models produce consistently lower information over the central portion of the theta scale. This is in line previous work suggesting that unidimensional models over-estimate information in the presence of LID.

With the presence of testlets established, discussion of next steps are in order. Some next steps involve more intensive psychometric studies, while others involve more practical analyses which can be carried out rather quickly to view the impact of these findings on test use.

Given the differences seen between the unidimensional model and the TRT model used here, it is reasonable to explore how other higher-order or multidimensional models may work in estimating parameters from these data. As the TRT models constrain the a parameter to be equal between the ability estimate and the testlet-specific gamma, it will

be of particular interest to explore what happens when the a parameter is free to vary between these latent estimates. Thus, a range of models accounting for the LID in different ways can be used to estimate relevant parameters, which can subsequently be compared.

Secondly, marginal reliability should be compared between the traditional 2PL case and the TRT models. Wainer and Thissen (1996) showed that LID is present when differences arise between marginal reliability estimated from unidimensional and testlet models (Wainer & Thissen, 1996). Thus, calculating IRT-based marginal reliability with the standard 2PL model to that estimated from the testlet model not only give an additional measure of the amount of LID present, but also give a sense of the differences between the models. While calculating the marginal reliability from traditional models is straightforward (Green, Bock, Humphreys, Linn, & Reckase, 1984), calculating a marginalized version still requires additional attention in the literature, as the format suggested by Wainer and colleagues may not be the best estimate given the differences in interpretation of the a parameter in the TRT models (Ip, 2010). More study on marginal reliability in TRT and other multi-dimensional models will be useful to this test specifically, and also more generally will be of import to the field.

More practical analyses can be done rather quickly using existing data, given the results of this study. First, given the relatively high correlation and low differences between theta estimates from each model, would use of the TRT model change any of the cut-point pass decisions for actual examinees? Using the more complex model may make sense only to the extent to which it impacts the pass/fail decisions of actual examines. Parameters using this model could be used to re-estimate theta values, determine a pass

decision, and examine how many examinees are impacted in either direction from the use of a different model.

Additionally, the content and skill areas, along with format, can be compared from tasks with small levels of testlet variance to tasks with greater levels of testlet variance in order to determine whether the content area, the skill assessed, or the format used is associated with and can explain the levels of testlet effects. This information can be applied to the task of content development and form construction. If different types of tasks generate varying amounts of testlet effects, this can inform task development through intentional targeting of either tasks with more highly-related items or by targeting tasks with essentially uncorrelated items, depending on the test developer's needs and the results of the psychometric studies suggested above.

Together, these results show that the local item independence assumption using these data is violated. This knowledge can be used to explore alternative models to estimate parameters used in test scoring which can account for this violation. Ultimately, the decision to use any alternative model must be made after only first examining the impact of alternative model use on the pass/fail decisions of actual examinees. That data can in turn drive whether the local dependence from task-based simulations is something to be encouraged or avoided. If content developers can alter the amount of dependency by creating tasks with specific interactions between content, skill, and format, those item types with the amount of dependency desired can be encouraged and ultimately utilized in test form construction.

Table 1. Summary of forms, sample size (*N*), and number of task items by section and by testlet.

Section	Panel	*N*	Items	Independent Items	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6
1	1	811	36	2	6	6	6	6	6	4
1	2	814	30	1	4	7	2	4	7	5
1	3	874	32	2	4	6	7	4	4	5
2	1	707	32	2	6	3	4	6	6	5
2	2	671	43	2	6	6	8	8	7	6
2	3	719	34	2	8	5	7	6	4	2
3	1	740	33	2	6	4	6	5	5	5
3	2	750	37	2	8	5	4	6	6	6
3	3	771	37	2	7	7	7	4	6	4

Table 2. Average Q_3 values on all unique item pairs within each testlet estimated through BILOG-MG. In parentheses is the expected value, –1/(*n*-1), with *n* being the number of test items.

Section	Panel	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6
1	1	0.015	0.063	0.004	0.068	0.045	0.052
		(-0.200)	(-0.200)	(-0.200)	(-0.200)	(-0.200)	(-0.333)
1	2	0.071	0.007	-0.001	0.034	0.005	0.172
		(-0.333)	(-0.167)	(-1.000)	(-0.333)	(-0.167)	(-0.250)
1	3	0.076	0.053	0.031	0.091	0.117	0.156
		(-0.333)	(-0.200)	(-0.167)	(-0.333)	(-0.333)	(-0.250)
2	1	0.261	0.191	0.240	0.105	0.075	0.275
		(-0.200)	(-0.500)	(-0.333)	(-0.200)	(-0.200)	(-0.250)
2	2	0.199	0.201	0.156	0.104	0.083	0.187
		(-0.200)	(-0.200)	(-0.143)	(-0.143)	(-0.167)	(-0.200)
2	3	0.144	0.120	0.168	0.177	0.322	0.416
		(-0.143)	(-0.250)	(-0.167)	(-0.200)	(-0.333)	(-1.000)
3	1	0.109	0.139	0.091	0.106	0.059	0.041
		(-0.200)	(-0.333)	(-0.200)	(-0.250)	(-0.250)	(-0.250)
3	2	0.104	0.059	0.325	0.223	0.173	0.071
		(-0.143)	(-0.250)	(-0.333)	(-0.200)	(-0.200)	(-0.200)
3	3	0.156	0.072	0.007	0.109	0.241	0.151
		(-0.167)	(-0.167)	(-0.167)	(-0.333)	(-0.200)	(-0.333)

Table 3. Gelman-Rubin F-tests for convergence of item parameters for SCORIGHT 2PL and SCORIGHT 2PL Testlet models. Values closer to 1.00 are ideal, with values less than or equal to 1.20 acceptable.

Section	Panel	a	b	Var(a)	Var(b)	Cov(a,b)
2PL						
1	1	1.00	1.00	1.00	1.00	1.00
1	2	1.00	1.00	1.00	1.00	1.00
1	3	1.00	1.00	1.00	1.00	1.00
2	1	1.00	1.00	1.00	1.00	1.00
2	2	1.00	1.00	1.00	1.00	1.00
2	3	1.00	1.00	1.00	1.00	1.00
3	1	1.00	1.00	1.00	1.00	1.00
3	2	1.00	1.00	1.00	1.00	1.00
3	3	1.00	1.00	1.00	1.00	1.00
2PL Testlet						
1	1	1.00	1.00	1.00	1.01	1.01
1	2	1.00	1.00	1.00	1.00	1.00
1	3	1.00	1.00	1.00	1.00	1.00
2	1	1.00	1.00	1.00	1.00	1.00
2	2	1.00	1.00	1.00	1.00	1.00
2	3	1.00	1.00	1.00	1.00	1.00
3	1	1.00	1.00	1.00	1.00	1.00
3	2	1.00	1.00	1.00	1.00	1.00
3	3	1.01	1.00	1.00	1.00	1.00

Table 4. Gelman-Rubin F-tests for convergence of testlet variance estimates for SCORIGHT 2PL Testlet models. Values closer to 1.00 are ideal, with values less than or equal to 1.20 acceptable.

Section	Panel	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6
1	1	1.01	1.03	1.04	1.01	1.01	1.01
1	2	1.02	1.01	1.06	1.02	1.04	1.07
1	3	1.01	1.01	1.01	1.01	1.02	1.02
2	1	1.00	1.02	1.01	1.00	1.00	1.02
2	2	1.01	1.03	1.01	1.01	1.01	1.02
2	3	1.01	1.01	1.01	1.01	1.03	1.03
3	1	1.01	1.01	1.01	1.01	1.01	1.02
3	2	1.01	1.01	1.01	1.01	1.04	1.02
3	3	1.02	1.03	1.05	1.02	1.00	1.02

Table 5. Average Q_3 values on all unique item pairs within each testlet estimated with Testlet Response Models in SCORIGHT. In parentheses is the expected value, –1/(*n*-1), with *n* being the number of test items.

Section	Panel	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6
1	1	-0.141	0.150	-0.188	-0.054	0.185	-0.123
		(-0.200)	(-0.200)	(-0.200)	(-0.200)	(-0.200)	(-0.333)
1	2	-0.241	-0.147	-0.577	-0.182	-0.139	0.177
		(-0.333)	(-0.167)	(-1.000)	(-0.333)	(-0.167)	(-0.250)
1	3	-0.156	-0.061	-0.085	-0.175	-0.147	0.172
		(-0.333)	(-0.200)	(-0.167)	(-0.333)	(-0.333)	(-0.250)
2	1	0.296	-0.055	0.143	0.071	-0.006	0.228
		(-0.200)	(-0.500)	(-0.333)	(-0.200)	(-0.200)	(-0.250)
2	2	0.191	0.230	0.237	0.185	0.189	0.132
		(-0.200)	(-0.200)	(-0.143)	(-0.143)	(-0.167)	(-0.200)
2	3	0.149	0.023	0.129	0.112	0.160	-0.316
		(-0.143)	(-0.250)	(-0.167)	(-0.200)	(-0.333)	(-1.000)
3	1	0.012	0.001	0.131	0.110	-0.190	-0.211
		(-0.200)	(-0.333)	(-0.200)	(-0.250)	(-0.250)	(-0.250)
3	2	0.172	-0.204	0.102	0.167	0.157	-0.026
		(-0.143)	(-0.250)	(-0.333)	(-0.200)	(-0.200)	(-0.200)
3	3	0.263	0.040	-0.160	0.053	0.153	0.144
		(-0.167)	(-0.167)	(-0.167)	(-0.333)	(-0.200)	(-0.333)

Table 6. Testlet gamma (γ) variances by section and panel.

Section	Panel	Testlet 1	Testlet 2	Testlet 3	Testlet 4	Testlet 5	Testlet 6
1	1	0.576	2.567	0.755	1.374	2.532	4.627
1	2	0.659	0.945	1.036	1.070	0.542	8.121
1	3	1.516	0.903	1.406	1.219	1.452	8.173
2	1	1.962	2.101	1.308	0.814	0.737	2.547
2	2	1.527	1.975	1.266	4.106	4.362	0.986
2	3	1.700	1.295	1.636	2.107	3.046	2.001
3	1	3.300	2.405	1.712	1.543	1.323	0.649
3	2	1.924	0.520	9.347	6.452	3.033	1.464
3	3	2.504	0.784	0.317	3.783	7.648	2.728

Table 7. Summary descriptive statistics for a, b, and θ values for all models.

Parameter descriptives are in the form average±SD (range).

Section	Panel	Items	N	a	b	θ
2PL BILOG-MG						
1	1	36	811	0.436±0.053	-0.519±1.518	0.000±0.877
				(0.317-0.571)	(-3.093-2.857)	(-3.548-2.262)
1	2	30	814	0.402±0.041	-0.461±1.602	0.000±0.842
				(0.328-0.521)	(-4.050-2.945)	(-2.736-2.089)
1	3	32	874	0.429±0.042	-0.783±1.344	0.000±0.862
				(0.354-0.523)	(-3.014-3.487)	(-3.183-2.305)
2	1	32	707	0.658±0.101	0.742±1.133	0.003±0.933
				(0.454-0.873)	(-1.327-3.094)	(-2.317-2.295)
2	2	43	671	0.612±0.089	0.622±0.993	0.000±0.940
				(0.467-0.782)	(-0.778-2.866)	(-2.563-2.954)
2	3	34	719	0.591±0.059	-0.061±0.914	0.000±0.928
				(0.485-0.701)	(-1.695-2.405)	(-2.835-2.809)
3	1	33	740	0.461±0.058	0.855±1.171	0.001±0.877
				(0.359-0.604)	(-1.250-3.418)	(-2.471-2.410)
3	2	37	750	0.488±0.040	0.305±1.682	-0.001±0.892
				(0.409-0.572)	(-2.772-3.194)	(-3.189-2.659)
3	3	37	771	0.517±0.065	0.709±1.721	0.000±0.903
				(0.402-0.659)	(-2.400-3.565)	(-2.958-2.232)
2PL SCORIGHT						
1	1	36	811	0.444±0.156	-0.519±1.518	0.000±0.877
				(0.146-0.874)	(-3.130-3.196)	(-3.521-2.301)
1	2	30	814	0.418±0.122	-0.461±1.602	0.000±0.842
				(0.205-0.758)	(-3.325-3.191)	(-2.694-2.128)
1	3	32	874	0.428±0.111	-0.783±1.344	0.000±0.862
				(0.239-0.709)	(-2.784-3.173)	(-3.241-2.230)
2	1	32	707	0.782±0.407	0.742±1.133	0.003±0.933
				(0.274-1.816)	(-1.184-3.375)	(-2.279-2.430)
2	2	43	671	0.648±0.247	0.622±0.993	0.000±0.940
				(0.285-1.309)	(-1.083-3.209)	(-2.574-2.907)
2	3	34	719	0.598±0.118	-0.061±0.914	0.000±0.928
				(0.394-0.845)	(-1.518-2.224)	(-2.842-2.827)
3	1	33	740	0.469±0.186	0.855±1.171	0.001±0.877
				(0.193-1.008)	(-1.488-3.184)	(-2.413-2.442)
3	2	37	750	0.487±0.088	0.305±1.682	-0.001±0.892
				(0.343-0.725)	(-2.497-3.227)	(-3.179-2.640)
3	3	37	771	0.578±0.211	0.709±1.721	0.000±0.903
				(0.294-1.016)	(-1.886-3.896)	(-2.870-2.228)
2PL SCORIGHT Testlet						
1	1	36	811	0.695±0.356	-0.519±1.518	0.000±0.877
				(0.099-1.330)	(-4.907-3.266)	(-3.479-2.119)

1	2	30	814	0.617±0.293	-0.461±1.602	0.000±0.842
				(0.193-1.420)	(-4.971-2.853)	(-2.767-2.065)
1	3	32	874	0.610±0.230	-0.783±1.344	0.000±0.862
				(0.196-0.994)	(-4.672-2.550)	(-3.144-2.335)
2	1	32	707	0.847±0.318	0.742±1.133	0.003±0.933
				(0.320-1.700)	(-1.138-3.360)	(-2.253-2.434)
2	2	43	671	1.355±0.763	0.622±0.993	0.000±0.940
				(0.271-2.833)	(-0.791-3.730)	(-2.400-2.940)
2	3	34	719	0.806±0.267	-0.061±0.914	0.000±0.928
				(0.225-1.266)	(-1.492-2.186)	(-2.758-2.726)
3	1	33	740	0.553±0.234	0.855±1.171	0.001±0.877
				(0.258-1.030)	(-1.463-3.060)	(-2.382-2.399)
3	2	37	750	0.660±0.231	0.305±1.682	-0.001±0.892
				(0.272-1.056)	(-2.295-4. 959)	(-3.081-2.573)
3	3	37	771	0.849±0.366	0.709±1.721	0.000±0.903
				(0.287-1.532)	(-1.681-4.542)	(-2.755-2.196)

Table 8. Pearson correlations between BILOG-MG parameter estimates and estimates from the SCORIGHT 2PL and SCORIGHT 2PL testlet models.

Section	Panel	SCORIGHT			SCORIGHT Testlet		
		a	b	θ	a	b	θ
1	1	0.987	0.988	0.993	0.630	0.907	0.970
1	2	0.965	0.983	0.996	0.833	0.925	0.972
1	3	0.987	0.990	0.996	0.718	0.883	0.979
2	1	0.947	0.988	0.982	0.755	0.986	0.973
2	2	0.969	0.981	0.992	0.663	0.883	0.958
2	3	0.993	0.997	0.998	0.709	0.970	0.977
3	1	0.970	0.980	0.990	0.734	0.958	0.980
3	2	0.978	0.997	0.999	0.718	0.955	0.963
3	3	0.967	0.987	0.993	0.748	0.929	0.970

Table 9. Root Mean Square Difference (RMSD) and Average Signed Difference (ASD) between BILOG-MG parameter estimates and estimates from the SCORIGHT 2PL and SCORIGHT 2PL testlet models.

	RMSD		ASD	
	SCORIGHT	**SCORIGHT Testlet**	**SCORIGHT**	**SCORIGHT Testlet**
Section 1, Panel 1				
a	0.103	0.412	-0.007	-0.258
b	0.237	0.647	0.000	0.000
θ	0.100	0.215	0.000	0.000
Section 1, Panel 2				
a	0.084	0.334	-0.016	-0.216
b	0.291	0.612	0.000	0.000
θ	0.076	0.200	0.000	0.000
Section 1, Panel 3				
a	0.068	0.269	0.001	-0.182
b	0.184	0.638	0.000	0.000
θ	0.074	0.177	0.000	0.000
Section 2, Panel 1				
a	0.332	0.311	-0.124	-0.189
b	0.170	0.186	0.000	0.000
θ	0.176	0.216	0.000	0.000
Section 2, Panel 2				
a	0.165	1.020	-0.036	-0.743
b	0.192	0.474	0.000	0.000
θ	0.119	0.271	0.000	0.000
Section 2, Panel 3				
a	0.060	0.311	-0.007	-0.215
b	0.075	0.222	0.000	0.000
θ	0.065	0.198	0.000	0.000
Section 3, Panel 1				
a	0.130	0.214	-0.009	-0.092
b	0.231	0.336	0.000	0.000
θ	0.123	0.176	0.000	0.000
Section 3, Panel 2				
a	0.049	0.265	0.001	-0.172
b	0.124	0.500	0.000	0.000
θ	0.042	0.241	0.000	0.000
Section 3, Panel 3				
a	0.159	0.458	-0.061	-0.332
b	0.278	0.639	0.000	0.000
θ	0.105	0.223	0.000	0.000

Figure 1. Relative efficiencies at each point of theta, Information(TRT) / Information(2PL) for all panels of Section 1. Panel 1 is represented by the blue curve, panel 2 by the purple curve, and panel 3 by the green curve.

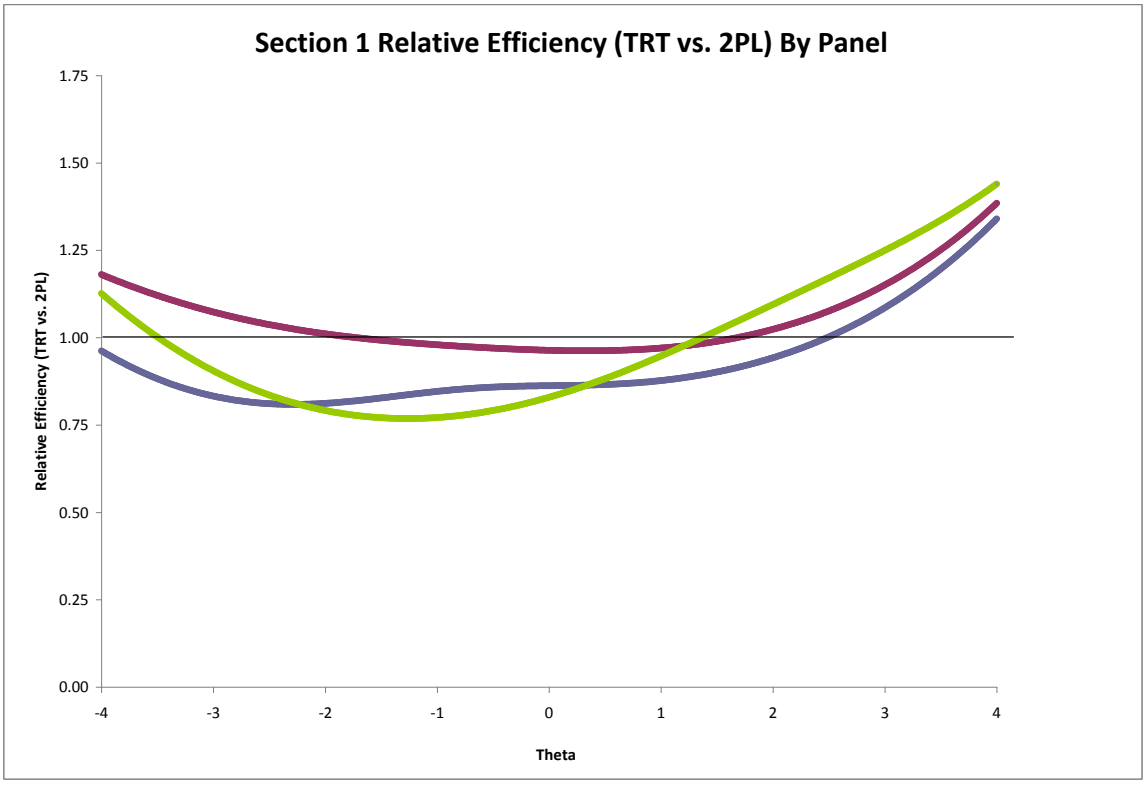


Figure 2. Relative efficiencies at each point of theta, Information(TRT) / Information(2PL) for all panels of Section 2. Panel 1 is represented by the blue curve, panel 2 by the purple curve, and panel 3 by the green curve.

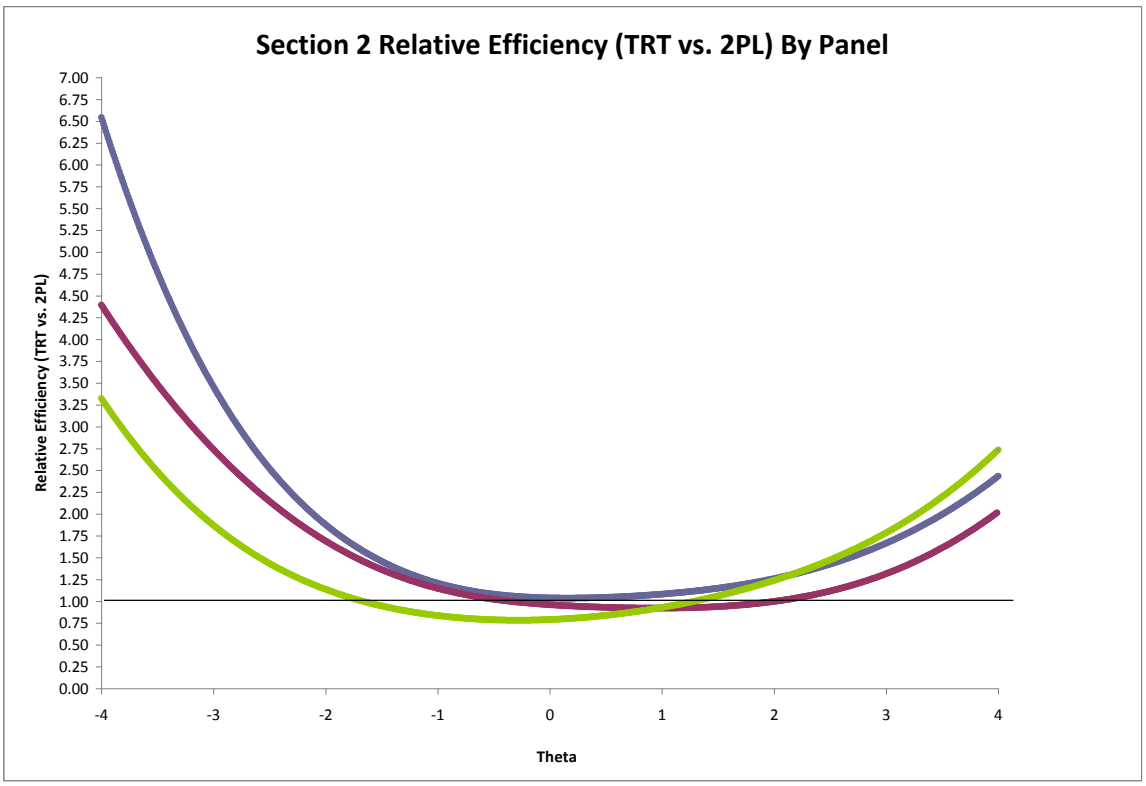
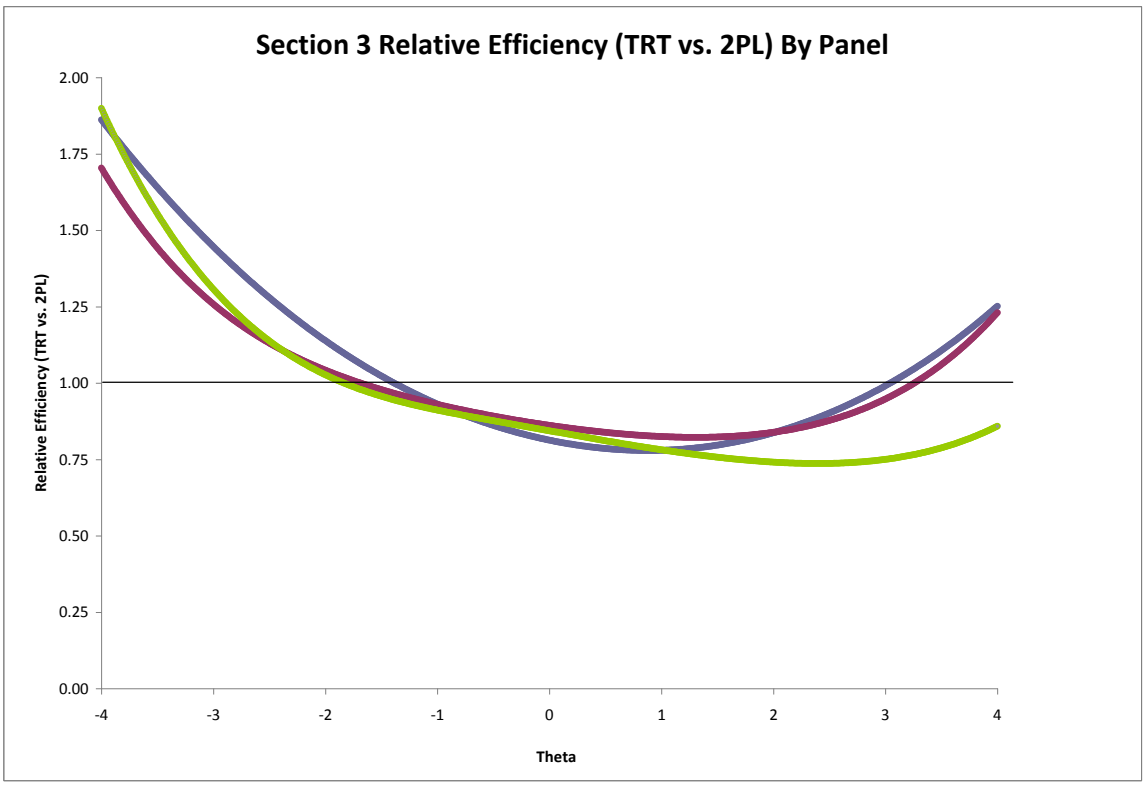


Figure 3. Relative efficiencies at each point of theta, Information(TRT) / Information(2PL) for all panels of Section 3. Panel 1 is represented by the blue curve, panel 2 by the purple curve, and panel 3 by the green curve.



References.

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* Paper presented at the Annual meeting of the American Educational Research Association, Washington, DC.

Birnbaum, A. (1968). Some latent trait scores and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

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

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

de la Torre, & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model Approach. *Applied Psychological Measurement, 33*(8), 620-639.

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

Du, Z. (1998). Modeling conditional item dependencies with a three-parameter logistic testlet model. *Dissertation Abstracts International, 59*(10), 5429.

Gelman, A., & Rubin, D. (1993). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472.

Green, B., Bock, R., Humphreys, L., Linn, R., & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*(21), 347-360.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, NJ: Sage.

Ip, E. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*, 467-482.

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement for testlet-based test scores using simulation techniques. *Journal of Educational Measurement, 36*(2), 91-112.

Lee, G., Kolen, M., Frisbie, D., & Ankenmann, R. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*(4), 357–372.

Lee, Y. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*(1), 74-100.

Li, Y., Bolt, D., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3–21.

Li, Y., & Cohen, A. (2003). *Equating tests composed of testlets: A comparison of a testlet response model and four polytomous response models*. Paper presented at the Annual meeting of the National Council on Measurement in Education.

Mislevy, R., & Bock, R. (1991). BILOG user's guide. Chicago, IL: Scientific Software.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114.

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

Stark, S., Chernyshenko, O., & Orleans., F. D. N. (2002). *Estimating the effects of local dependence on the accuracy of IRT estimation.* Paper presented at the National Council on Measurement in Education, New Orleans, LA.

Thissen, D. (1993). MULTILOG user's guide Version 6.3. Mooresville IN: Scientific Software.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiplecategorical response models. *Journal of Educational Measurement, 26*(3), 247-260.

Wainer, H. (1995). Precision & differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*(2), 157-187.

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in adaptive testing. In W. v. d. Linden & C. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: The case for testlets. *Journal of Educational Measurement, 24*, 189-205.

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*(5), 749-766.

Wainer, H., & 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 Practices, 15*(1), 22-29.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Wang, X., Bradlow, E., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*(1), 109-128.

Wang, X., Bradlow, E., & Wainer, H. (2004). User's guide for SCORIGHT, version 3.1: a computer program for scoring tests built of testlets. Princeton, NJ: Educational Testing Service.

Yen, W. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 8*, 125-145.

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

Yung, Y., Thissen, D., & McLeod, L. (1999). On the relationship between the higher-order model and the hierarchical factor model. *Psychometrika, 64*, 113-128.