Measuring the Power of Learning.®

Research Report
ETS RR–17-27

An Information-Correction Method for Testlet-Based Test Analysis: From the Perspectives of Item Response Theory and Generalizability Theory

Feifei Li

June 2017

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

An Information-Correction Method for Testlet-Based Test Analysis: From the Perspectives of Item Response Theory and Generalizability Theory

Feifei Li

Educational Testing Service, Princeton, NJ

An information-correction method for testlet-based tests is introduced. This method takes advantage of both generalizability theory (GT) and item response theory (IRT). The measurement error for the examinee proficiency parameter is often underestimated when a unidimensional conditional-independence IRT model is specified for a testlet dataset. By using a design effect ratio composed of random variances that can be easily derived from GT analysis, it becomes possible to adjust the underestimated measurement error from the unidimensional IRT models to a more appropriate level. In this paper, it is demonstrated how the information-correction method can be implemented in the context of a testlet design. Also, through the simulation study, it is shown that the underestimated measurement errors from IRT estimates can be adjusted to the appropriate level despite the varying magnitudes of local item dependence (LID), testlet length, balance of testlet length, and number of item parameters in the model. The real data example provides more details about when and how the information-correction method should be used in a test analysis. Estimation by the information-correction method should be adequate for practical work, given the robustness of the variance ratio.

Keywords Item response theory; generalizability theory; testlet, information-correction method; local item dependence

doi:10.1002/ets2.12151

Testlet indicates a set of items sharing a single common stimulus (Rosenbaum, 1988) where the performance on each item depends on both a general ability and a specific ability related to the particular content or occasion — for example, a reading passage or an information graph. Testlets help to develop a more realistic and contextualized test. It is also thought that testlets integrate the knowledge and skills that cannot be represented in simple independent multiple-choice items. They can provide insights into not only general abilities, but also a series of specific cognitive information processing in complex tasks (Rosenbaum, 1988; Sternberg, 1977). With testlets, the time and cost for collecting additional information can be reduced. As testlets can bring beneficial consequences to educational practices (Messick, 1994), they have been seen in many large-scale tests.

However, due to the particular statistical properties of testlets, issues have emerged in regard to applying unidimensional measurement models to the testlet datasets. One of the properties that has brought up many technical concerns is *local item dependence* (LID) — namely, the common stimulus that the set of items rely upon can introduce dependence among the responses within an individual. For example, when some students have a special interest or better prior background knowledge in a passage than other students, they are likely to perform better on the items related to this passage than on other items of the same difficulty level, or they tend to perform better than other students with the same general ability level.

In contrast, conditional independence (CI) or local item independence is assumed in the conventional item response theory (IRT) models. The CI assumption states that given the fixed ability level, an examinee's performance on one item must not affect his or her responses to any other items in the test. Unidimensional IRT models may not be robust to the violation of the CI assumption (Hambleton & Swaminathan, 1985). In that case, analyzing testlet datasets with misspecified unidimensional IRT models could lead to undesired results. As shown in a number of previous studies, ignoring LID resulted in overestimated precision of ability estimates, and the problems of estimation were exacerbated when either the testlet length or the testlet effect increased (Bradlow, Wainer, & Wang, 1999; Sireci, Wainer, & Thissen, 1991; Wainer, 1995; Yen, 1993).

Corresponding author: Feifei Li, E-mail: fli@ets.org

The accuracy of ability estimates is particularly crucial under some circumstances. For example, if the test results are used in ways that have consequences for individual examinees, greater accuracy is required at the score level. When the cut scores are applied to proficiency classification, the measurement errors of proficiency estimates also need to be considered. For another example, in computer adaptive testing, overestimation of precision would present difficulties for setting stopping rules and lead to premature termination (Du, 1998; Wainer, Bradlow, & Du, 2000).

To account for LID from the response patterns of testlets, a number of nonparametric and parametric approaches have been created and employed. Generalizability theory (GT) has been traditionally used to model and analyze a variety of statistical dependencies on the raw score scales (Brennan, 1992; Cronbach, Linn, Brennan, & Haertel, 1997; Koretz, Stecher, Klein, & McCaffrey, 1994; Lee & Frisbie, 1999; Sireci et al., 1991). By using GT, one does not have to demonstrate the satisfaction of strong statistical assumptions that are required by IRT. A GT approach has been regarded as a convenient method, as it can easily partition the variances from different resources and provide the information about the reliabilities and errors of measurement. However, GT was originally created for continuous variables rather than for discrete item scores (Brennan, 1997). Although a hybrid approach that incorporates GT and IRT has been developed to fulfill a nonlinear transformation from the discrete raw test scores to the continuous item and person variables, this approach has currently been limited to the single-facet measurement design with binary items (Briggs & Wilson, 2007).

In contrast, IRT models specify a probabilistic relationship between the item responses and the characteristics of the individuals and the items. The link function makes it possible to connect the discrete responses with the continuous latent variables. Testlet models from the IRT approach generally capture the person–testlet interactions in terms of multidimensional variables modeled as random effects—for example, the Rasch testlet and random-effects facet models (Wang & Wilson, 2005a, 2005b), which are special cases of the multidimensional random coefficients multinomial logit model (MRCMLM) by Adams, Wilson, and Wang (1997); the bifactor model (Gibbons & Hedeker, 1992); the multilevel model (Jiao, Wang, & Kamata, 2005); and testlet response theory (TRT) models (Bradlow et al., 1999; Wainer, Bradlow, & Wang, 2007; Wainer et al., 2000; Wang, Bradlow, & Wainer, 2002). As has been shown in a series of simulation and real data studies (Bradlow et al., 1999; DeMars, 2006; Jiao & Wang, 2008; Jiao et al., 2005; Wainer et al., 2000, 2007; Wang & Wilson, 2005a, 2005b; Wang et al., 2002), these testlet IRT models demonstrate good model fit, small bias, and satisfactory accuracy in parameter recovery on the testlet datasets compared with their unidimensional IRT counterparts.

The feasibility of the estimation of these testlet response models depends to a large extent on the recent increase in computational power. Marginal maximum likelihood estimation (MMLE) with the expectation-maximization (EM) algorithm has been applied in MRCMLM and the bifactor model; penalized quasilikelihood estimation or Laplace approximation (Laplace) has been often used for the multilevel models; Markov Chain Monte Carlo (MCMC) is the method for estimation in TRT models. In comparison, Laplace and MCMC yielded accurate parameter recovery and appropriate precision of estimates but took a very long time to converge (Jiao & Wang, 2008; Sinharay, 2003) and thus have been rarely implemented in operational testing. MMLE with the EM algorithm was relatively efficient, and its performance in parameter estimation was adequate (Demars, 2006; Jiao & Wang, 2008). However, in these applications of MMLE, ability and testlet parameters were estimated conditional on the point estimates of the item parameters, so the uncertainty in the estimation of the item parameters has been ignored (Wainer et al., 2007). In addition, it is noteworthy that adaptive quadrature has a quite major effect on computations with maximum marginal likelihood (Haberman, 2013).

Considering the design of testlets where items are clustered versus the design of the independent items, the downward-biased estimation of the variance of ability estimates as a result of misspecifying the unidimensional IRT models on the testlet data may be adjusted through the design effect. Bock, Brennan, and Muraki (2002) proposed correcting the information function of multiple ratings by using a variance ratio term derived from the GT analysis. Considering the similarity between the testlets and multiple ratings in terms of local dependency between the responses in clusters, this method is extended to the situation of testlets to adjust the underestimated measurement error of abilities. The design effect is often used as a measure of the precision gained or lost by the use of a more complex design instead of simple random sampling, as discussed by Cornfield (1951). With respect to its estimation procedure, it is relatively efficient to obtain the design effect by deriving the variance of person estimates of either design through GT, and the information of the ability estimates in independent item design through maximum likelihood estimation (MLE). Hence, given their strengths, GT and IRT can jointly contribute to adjust the information of ability estimates in the testlet-based tests to a more appropriate level.

This paper provides both a description of the information-correction method and an evaluation of the approach. To achieve this purpose, it is necessary to (a) explore a computational approach for this information-correction method, (b) conduct a simulation study to evaluate the performance of the proposed information-correction method under conditions with varying factors, and (c) apply the information-correction method to one example with real data.

Theoretical Framework

The formula below represents the GT model for testlets. Assuming the testlet dataset has a univariate nested design of $i \times (j:d)$, that is, persons i crossed with j items nested in d testlet, the mean of the item scores in a given set of items is the best linear unbiased estimator of an individual, the linear model of which can be represented as follows assuming person is completely random (Lee & Frisbie, 1999):

$$\begin{aligned} X_{ij:d} = {} & \mu && \text{(grand mean)} \\ & + \mu_i - \mu && \text{(person effect)} \\ & + \mu_d - \mu && \text{(testlet effect)} \\ & + \mu_{j:d} - \mu_d && \text{(item within testlet effect)} \\ & + \mu_{id} - \mu_i - \mu_d + \mu && \text{(person} \times \text{testlet interaction effect)} \\ & + X_{ijd} - \mu_{id} - \mu_{j:d} + \mu_d && \text{(residual effect)} \end{aligned} \tag{1}$$

In the GT model of testlets, the magnitude of the testlet effect is measured by variance of person-by-testlet interaction $\sigma^2(iJ{:}D)$.

To construct a correction for estimation of accuracy of proficiency estimates due to use of testlets, begin with a simple situation where the main effects of items and testlets are assumed to be fixed. In that case, among all the partitioned effect terms previously noted (Equation 1), the person effect, the person-by-testlet interaction, and the person-by-item within-testlet confounded with residuals are random effects to be involved in the generalizability analysis of the current study. It happens that, in the linear predictors of either IRT or TRT models, they are also regarded as facets with random effects, but the item difficulties are fixed effects.

In the GT framework, the total variance can be expressed as $\sigma^2(X) = \sigma^2(\tau) + \sigma^2(\delta)$, in which $\sigma^2(\tau)$ represents the variance of the universe score that is based only on the variance of the person covariates, $\sigma^2(i)$, but the relative error variance is composed of variance of the person-by-testlet interaction $\sigma^2(iD)$ and variance of persons by items-within-testlet interaction $\sigma^2(iJ{:}D)$, which can be expressed as $\sigma^2(\delta^2) = \sigma^2(iD) + \sigma^2(iJ{:}D)$.

The proficiency of examinee i can be estimated by the mean scores across all items and all testlets,

$$X_{ij:d} = \tau_i + \varepsilon_{id} + \xi_{ij:d}, \tag{2}$$

where $j = 1, \ldots, n$, $d = 1, \ldots m$, τ_i denotes the grand mean, ε_{id} denotes the person-by-testlet interaction effect, $\xi_{ij:d}$ denotes the person-by-item within-testlet interaction effect, $\sigma^2(i)$, $\sigma^2(iD)$, and $\sigma^2(iJ{:}D)$ are variance terms on τ_i, ε_{id}, and $\xi_{ij:d}$, respectively. The total number of items is calculated by $n = \sum_{d=1}^{m} k_d$, where k_d indicates the number of items in each testlet. When the test has a balanced design with equal number of items in each testlet, $k_d = k = n/m$. The mean score of each examinee is

$$\begin{aligned} X_{i..} &= \frac{1}{n} \sum_{d}^{m} \sum_{j:d}^{k_d} X_{ij:d} \\ &= \tau_i + \frac{1}{m} \sum_{d}^{m} \varepsilon_{id} + \frac{1}{n} \sum_{d}^{m} \sum_{j:d}^{k_d} \xi_{ij:d}. \end{aligned} \tag{3}$$

The variance of the mean score estimate is

$$\sigma^2\left(\mathrm{X}_{i..}\right) = \sigma^2\left(\mathrm{i}\right) + \frac{1}{m}\sigma^2\left(\mathrm{iD}\right) + \frac{1}{\mathrm{n}}\sigma^2\left(\mathrm{iJ} : \mathrm{D}\right). \tag{4}$$

The random error variance component is

$$\frac{1}{m}\sigma^2\,(\mathrm{iD}) + \frac{1}{\mathrm{n}}\sigma^2\,(\mathrm{iJ}:\mathrm{D})\,. \tag{5}$$

Suppose the testlet facet is ignored as if all items were independent. The variance other than the true variance is the error variance, which in this case is composed of the variance of the interaction between persons and items.

$$\begin{aligned} X_{i.} &= \frac{1}{n}\sum_{j}^{n} X_{ij} \\ &= \tau_i + \frac{1}{n}\sum_{d}^{m}\sum_{j:d}^{k_d}(\varepsilon_{id} + \xi_{ij:d}). \end{aligned} \tag{6}$$

The variance of the mean score estimate should be

$$\begin{aligned} \sigma^2\left(\mathrm{X}_{i.}\right) &= \sigma^2\,(\mathrm{i}) + \frac{1}{n}\sigma^2\,(\mathrm{iJ}) \\ &= \sigma^2\,(\mathrm{i}) + \frac{1}{n}\left[\sigma^2\,(\mathrm{iD}) + \sigma^2\,(\mathrm{iJ}:\mathrm{D})\right]. \end{aligned} \tag{7}$$

The random error variance is

$$\frac{1}{n}\left[\sigma^2\,(\mathrm{iD}) + \sigma^2\,(\mathrm{iJ}:\mathrm{D})\right]. \tag{8}$$

The ratio of the random error variances under the testlet design (Equation 5) and the independent item design (Equation 8),

$$\frac{\sigma^2\,(iD)/m + \sigma^2\,(iJ:D)/n}{\left[\sigma^2\,(iD) + \sigma^2\,(iJ:D)\right]/n}, \tag{9}$$

is used as a practical approximation term to correct the standard error variance when the testlet data are treated as independent responses and fit with conventional IRT models.

In the context of IRT, when a test consists of a sufficient number of items, the information function is asymptotically the reciprocal of the standard error variance of the MLE of ability estimates. Thus, the ratio to correct the testlet-specific information is

$$t_d = \frac{\sigma^2\,(iD) + \sigma^2\,(iJ:D)}{k_d\sigma^2\,(iD) + \sigma^2\,(iJ:D)}, \tag{10}$$

where k_{d} is the number of items in each testlet.

In a conventional IRT model that can be generalized to the categorical responses, the likelihood function for the $(N \times n)$ vector u of the responses of N examinees on n items is

$$\begin{aligned} L\,(\mathbf{u}|\boldsymbol{\theta}) &\equiv L\left(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N|\theta_1, \theta_2, \ldots, \theta_N\right) \\ &= \prod_{i=1}^{N} L(\mathbf{u}_i|\theta_i) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{n} P\left(u_{ij}|\theta\right) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{n}\prod_{h}^{q_j}\left[P_{hij}\,(\theta)\right]^{x_{hij}}. \end{aligned} \tag{11}$$

The logarithm of the likelihood function is

$$\ln L = \sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{h}^{q_j} x_{hij}\ln[P_{hij}\,(\theta)], \tag{12}$$

where i is the index of examinee, $(i = 1, \ldots, N)$; j is the index of items; $\boldsymbol{\theta}$ denotes the ability of examinees; $\boldsymbol{u}$ is the response vector consisting of x_{hij}; x_{hij} is the indicator variable taking on the value 1 if the response of examinee i to item j is assigned

Table 1 Analogous T Terms for Unbalanced *i**(*j*:*d*)

Source of variation	T
Examinees	$\mathrm{TE} = n\sum_i \overline{X}_i^2$
Testlets	$\mathrm{TT} = N\sum_d \left(n_{j:d}\overline{X}_d^2\right)$
Items: testlets	$\mathrm{TIT} = N\sum_d \sum_{j:d} \overline{X}_{j:d}^2$
Testlets × examinees	$\mathrm{TTE} = \sum_d \left(n_{j:d}\sum_i \overline{X}_{id}^2\right)$
(Items: testlets) × examinees	$\mathrm{TITE} = \sum_i \sum_j X_{ij}^2$
Mean (μ)	$\mathrm{T}_{\mathrm{mean}} n\overline{X}^2$

Note. $\overline{X}_i$ is the mean across all items for each person; $\overline{X}_d$ is the mean across a cluster of items and all persons for each testlet; $\overline{X}_{j:d}$ is the mean across all persons for each item; $\overline{X}_{id}$ is the mean across a cluster of items for each testlet and each person; $\overline{X}$ is the grand mean across all items and all persons; $n_{j:d}$ is the number of examines who complete item j nested in testlet *d*.

Table 2 Sum of Squares

Source of variation	Sum of squares
Examinees	$\mathrm{SSE} = \mathrm{TE} - \mathrm{T}_{\mathrm{mean}}$
Testlets × examinees	$\mathrm{SSTE} = \mathrm{TTE} - \mathrm{TT} - \mathrm{TE} - \mathrm{T}_{\mathrm{mean}}$
(Items: testlets) × examinees	$\mathrm{SSITE} = \mathrm{TITE} - \mathrm{TIT} - \mathrm{TTE} + \mathrm{TT}$

Table 3 Expected Random-Effect Variances

Source of variation	Examinees	Testlets × examinees	(Items: testlets) × examinees
Mean squares	$\mathrm{MSE} = \mathrm{SSE}/(N-1)$	$\mathrm{MSTE} = \mathrm{SSTE}/(N-1)(m-1)$	$\mathrm{MSITE} = \mathrm{SSI:TE}/(N-1)(n-m)$
Expected mean squares	$n\sigma^2(i) + r_d\sigma^2(iD) + \sigma^2(iJ:D)$	$w_d\sigma^2(iD) + \sigma^2(iJ:D)$	$\sigma^2(iJ:D)$
Estimated variance components	$\sigma^2(i) = [\mathrm{MSE} - r_d\mathrm{MSTE}/w_d + (r_d - w_d)\mathrm{MSITE}/t_d]/n$	$\sigma^2(iD) = (\mathrm{MSIE} - \mathrm{MSITE})/w_d$	$\sigma^2(iJ:D) = \mathrm{MSITE}$

to category *h*, and the value 0 otherwise, ($h = 1, \ldots, q_j$); P_{hij} is the categorical response probability for examinee *i* and item *j*.

For a testlet dataset where item *j* is nested in testlet *d*, a corrected quasilikelihood to estimate θ is

$$L_i^* = \prod_d^m \left\{ \prod_{j:d}^{k_d} \prod_h^{q_j} \left[P_{hij:d}(\theta)\right]^{x_{hij:d}} \right\}^{t_d}, \tag{13}$$

where *j*:*d* is the index of items in testlet *d* ($j = 1, \ldots, k_d$); k_d indicates the number of items in each testlet; *m* indicates the number of testlets; *d* is the index of the testlets; t_d is the ratio to correct the testlet-specific information; other symbols have the same meanings as in Equation 12. The rationale of t_d is explained above in Equation 1 through 10. It is assumed that the necessary variance components are estimated in a preliminary analysis of variance as shown in Tables 1 through 3. The log likelihood function is

$$\ln L_i^* = \sum_d^m t_d \sum_{j:d}^{k_d} \sum_h^{q_j} x_{hij:d} \cdot \ln\left[P_{hij:d}(\theta)\right]. \tag{14}$$

The first order condition to estimate θ is

$$\sum_d^m t_d \sum_{j:d}^{k_d} \sum_h^{q_j} \frac{x_{hij:d}}{P_{hij:d}(\theta)} \cdot \frac{\partial P_{hij:d}(\theta)}{\partial \theta} = 0, \tag{15}$$

Table 4 Simulation Design

Level of balance	Models	# Items per testlet	Variance of testlet effect ($\sigma^2_{\gamma h(i)}$)		
			.01	.25	1
Balanced	1PL	5	S1	S2	S3
		10	S4	S5	S6
	2PL	5	S7	S8	S9
		10	S10	S11	S12
	3PL	5	S13	S14	S15
		10	S16	S17	S18
Medium unbalanced	1PL	4, 6	S19	S20	S21
		8, 12	S22	S23	S24
	2PL	4, 6	S25	S26	S27
		8, 12	S28	S29	S30
	3PL	4, 6	S31	S32	S33
		8, 12	S34	S35	S36
Extreme unbalanced	1PL	2, 8	S37	S38	S39
		4, 16	S40	S41	S42
	2PL	2, 8	S43	S44	S45
		4, 16	S46	S47	S48
	3PL	2, 8	S49	S50	S51
		4, 16	S52	S53	S54

where $P_{hij:d}(\theta)$ is the category response probability for item j nested in testlet d.

The test information function is given by the following expression:

$$I_i(\theta) = -E\left\{\sum_d^m t_d \sum_{j:d}^{k_d}\sum_h^{q_j}\left\{\frac{1}{P_{hij:d}(\theta)}\left[\frac{\partial P_{hij:d}(\theta)}{\partial\theta}\right]^2\right\}\right\}. \tag{16}$$

For dichotomous items, $q_j = 2$, $P_{2j:d}(\theta) = 1\text{-}P_{1j:d}(\theta)$, $x_{2ij:d} = 1\text{-}x_{1ij:d}$, and the likelihood function is

$$\sum_d^m t_d \sum_{j:d}^{k_d}\left[P_{ij:d}(\theta)\right]^{x_{ij:d}}\left[1 - P_{ij:d}(\theta)\right]^{1-x_{ij:d}}, \tag{17}$$

where $P_{ij:d}$ is the response function for the correct response.

The following steps are implemented to obtain the correction from the generalizability analysis (shown in Table 1 through Table 3).

As presented in Table 4, the total number of examinees is N; the total number of items is n, and the number of testlets is m; the number of items in each testlet is k. For the balanced design, $r = w = k$; for the unbalanced design, $r_d = \sum_d \frac{k_d^2}{n}$ and $w = \frac{n-r}{m-1}$ (Brennan, 2001). The correction term for item-specific information is obtained through Equation 10.

In GT models, the mean of the set of item scores given to an examinee is the linear unbiased estimator of the ability of that individual, and the error variance is composed of random error variances of the facets and their interactions. TRT models produce relatively more accurate estimates, even from data with a significant magnitude of LID. Thus, we may conjecture that the variance of the ability parameter from the IRT model corresponds to the random error variance of estimates from an independent item design in GT, but the variance of the primary ability parameter from the TRT model corresponds to the random error variance of estimates from a testlet design in GT. Therefore, we can use the ratio of random error variances of the ability parameter in an independent design and a testlet design by generalizability analysis to adjust the estimated measurement error to a more appropriate level. This is the reasoning about the relationship between these error variances. However, it is necessary to further obtain the quantitative evidence in regard to the performance of the information-correction method through manipulating factors in simulation studies.

Methodology: Simulation Study

Design

The purpose of this simulation study is to evaluate the performance of the information-correction method in adjusting the measurement error of proficiency parameters by specifying the one-parameter logistic (1PL), the two-parameter logistic (2PL), and three-parameter logistic (3PL) IRT models to testlet datasets. The results are compared with the expected error variances from TRT models with the same number of item parameters. The research questions of interest in this study are as follows.

1. What factors have significant effects on the performance of the information-correction method? The performance of the method is evaluated by comparing the asymptotic standard error of estimates (SEEs) of proficiency from IRT models adjusted by the information-correction ratio and SEEs from TRT models. In this study, this criterion variable is named the standard error increase discrepancy (SEID) and formulated as

$$\frac{\mathrm{SEE}_{\mathrm{TRT}} - \mathrm{SEE}_{\mathrm{IRT}}}{\mathrm{SEE}_{\mathrm{IRT}}} - \frac{\mathrm{SEE}_{\mathrm{IRT}} t_d^{-1/2} - \mathrm{SEE}_{\mathrm{IRT}}}{\mathrm{SEE}_{\mathrm{IRT}}} = \frac{\mathrm{SEE}_{\mathrm{TRT}} - \mathrm{SEE}_{\mathrm{IRT}} t_d^{-1/2}}{\mathrm{SEE}_{\mathrm{IRT}}} \tag{18}$$

2. How do the distributional characteristics of the proficiency estimate change across the simulation conditions?

Previous studies have proposed that LID and testlet length are factors that might affect the parameter recovery from the testlet datasets as a result of model misspecification. Accuracy in proficiency and item parameter estimates deteriorates when either the testlet length or the magnitude of LID increases (Bradlow et al., 1999; Sireci et al., 1991; Wainer, 1995; Yen, 1993). In particular, the testlet models tend to provide more accurate and precise parameter estimates than the independent item models (Bradlow et al., 1999; DeMars, 2006; Jiao & Wang, 2008; Jiao et al., 2005; Wainer et al., 2000, 2007; Wang & Wilson, 2005a, 2005b; Wang et al., 2002). Modest testlet length tends to have minimal effect on precision of estimates if LID is ignored in the testlets (Bradlow et al., 1999; Wang et al., 2002). In addition, the information-correction ratio is a testlet-specific statistic that depends on the length of each testlet, so the adjusted SEE is weighted by the length of each testlet nonlinearly. Therefore, the balance of testlet length in the test also counts in this investigation in addition to LID and the test length.

The performance of the information-correction method needs to be investigated in the contexts of the 1PL, the 2PL, and the 3PL response theory models, respectively, because each of them has a different representation of the information function. For the 1PL model, because each item has the same discrimination value, the distributions of the information function are equal. For the 2PL model, each item has a different slope, and information functions are different from those of the 1PL model. The maximum amount of information provided by an item increases as the item discrimination increases. In the 3PL model, with the presence of guessing parameters, all other things equal, the amount of information an item provides decreases as the amount of guessing increases. These differences among the three types of models may lead to distinctive performances of the information-correction method.

Thus, three simulation factors are manipulated: (a) LID — the variance of the random testlet variables, specified at 0, .25, 1, representing zero, small, and large testlet effect respectively; (b) testlet length — short and long testlets (i.e., a testlet consisting of fewer than 10 items is regarded as the short testlet, whereas a testlet consisting of more than 10 items is regarded as the long testlet); and (c) balance of testlet length across the test-balanced design (i.e., equal number of items in each testlet), intermediately unbalanced design (e.g., 4 items in one testlet and 6 items in another or 8 items in one testlet and 12 items in another), and extremely unbalanced design (e.g., 2 items in one testlet and 8 items in another or 4 items in one testlet and 16 items in another). The data are generated by the 1PL, the 2PL, and the 3PL TRT models, respectively, and are calibrated using IRT or TRT models with the same number of item parameters. These three factors in the context of three models compose $3 \times 2 \times 3 \times 3 = 54$ conditions. The conditions are described and numbered in Table 4.

Data Generation

For each condition, the test consists of 60 dichotomous items, and the sample size of examinees is 500. Because the item parameters are known in the simulation and only the abilities need to be estimated, 500 is a sufficient sample size for this purpose. The probability of the correct response is calculated by using the 1PL TRT model (Equation 19), the 2PL TRT model (Equation 20), and the 3PL TRT model (Equation 21), respectively.

Table 5 Simulation Specifications

Parameters	Distributions
α	$\sim N(.8, .2)$
β	$\sim N(0, 1)$
ω	$\sim N(.14, .05)$
θ	$\sim N(0, 1)$
γ	Zero: =0 Small: $\sim N(0, .25)$ Large: $\sim N(0, 1)$

The probability of the correct response of the 1PL TRT model is

$$P\left(y_{ij}=1\right)=\frac{\exp\left(\theta_i-\beta_j-\gamma_{id(j)}\right)}{1+\exp\left(\theta_i-\beta_j-\gamma_{id(j)}\right)}, \tag{19}$$

where β_j is the item difficulty of the j_{th} item; θ_i is the person proficiency of the i_{th} person; $\gamma_{id(j)}$ parameterizes the random effect for person i on testlet d that contains item j; $P(y_{ij}=1)$ is the probability of correct response from person i on item j.

The probability of the correct response of the 2PL TRT model is

$$P\left(y_{ij}=1\right)=\frac{\exp\left[\alpha_j\left(\theta_i-\beta_j-\gamma_{id(j)}\right)\right]}{1+\exp\left[\alpha_j\left(\theta_i-\beta_j-\gamma_{id(j)}\right)\right]}, \tag{20}$$

where α_j is the item discrimination for item j.

The probability of the correct response of the 3PL TRT model is

$$P\left(y_{ij}=1\right)=\omega_j+\left(1-\omega_j\right)\frac{\exp\left[\alpha_j\left(\theta_i-\beta_j-\gamma_{id(j)}\right)\right]}{1+\exp\left[\alpha_j\left(\theta_i-\beta_j-\gamma_{id(j)}\right)\right]}, \tag{21}$$

where ω_j is the guessing parameter for item j.

The values of the parameters are generated from the distributions specified in Table 5. The set of true values of ability parameters is fixed across all conditions. The values of the testlet variable are randomly generated from a normal distribution with a mean of zero and a variance of $\sigma^2_{\gamma d(i)}$. The true values of the difficulty parameters (β) are generated from a standard normal distribution and truncated within −1.5 and 1.5. For responses from the 2PL and 3PL models, the true values of the discrimination parameters (α) are generated from a normal distribution truncated within the range of [.6, 1.4]. For responses from the 3PL model, the true values of the guessing parameters (ω) follow a normal distribution truncated within the range of [0, .25]. The marginal distributions of these parameters are chosen based on a typical form of achievement test. The item scores are simulated using the Bernoulli distribution function based on the probability of the correct response. Each condition is replicated 50 times.

Analysis

Estimation

In order to evaluate the performance of the information-correction method in adjusting the measurement error from IRT, each dataset is analyzed by IRT and TRT models, respectively. The models for estimation use the same number of item parameters as in the models for data generation. Among all the estimation methods, the Bayesian MCMC method provides a flexible and straightforward approach for estimating with either IRT or TRT models. Because the prior distribution represents the belief about the parameter and will pull the posterior estimate toward the prior mean, the incorporation of prior information will increase the meaningfulness and accuracy of the posterior estimates. The estimation with the MCMC method is implemented using WinBUGS embedded in R.

At the early stage of this research, because this study is targeted at the SEEs of the primary ability parameters, the item parameter values are fixed in order to speed up the estimation procedure. The prior distributions of the ability parameters

and random testlet variables are specified as follows:

$$\theta_i \sim N(0,1)$$
$$\gamma_{id(j)} \sim N\left(0, \sigma^2_{\gamma d(j)}\right). \tag{22}$$

The variance of the testlet effect $\sigma^2_{\gamma d(i)}$ is the random variance of examinee by items-within-testlet interaction and indicates the strength of LID for testlet $d(j)$. To estimate the variance of the testlet effect, a hyperprior for $\sigma^2_{\gamma d(i)}$ is specified as an inverse gamma distribution with shape parameter $\alpha = 1$ and scale parameter $\beta = 1$,

$$\sigma^2_{\gamma_{d(j)}} \sim \Gamma^{-1}(1,1). \tag{23}$$

To expedite convergence, ML point estimates of proficiency parameters from the IRT estimation are used as initial values. Two chains of iterations are run for each dataset. Convergence for a dataset of 60 items and 500 examinees usually occurs within 1,000 iterations (Bradlow et al., 1999; Wainer et al., 2000). To ensure that convergence would be achieved before a certain number of iterations, two chains of iterations are run first on a sample dataset generated using the same simulation specifications in each condition. Several convergence diagnostic criteria are available in WinBUGS: the dynamic trace lines, history plots, autocorrelation lines, Gelman – Rubin convergence statistics, and quantile plots.

The mean of the posterior distribution is regarded as the optimal estimate of the proficiency parameter, and the standard deviation of the posterior distribution is taken as the standard error of the proficiency estimate. Upon convergence, as one criterion ascertaining sufficient iterations have been run to best represent the posterior distribution, the MC errors should be no more than approximately 5% of the standard deviations of the posterior distributions.

The number of burn-in cycles and the sufficient number of iterations for the estimation of the posterior distributions depend on the complexity of the model and the sample size. For example, for estimation using the 2PL TRT model, 4,500 cycles are run for each chain and the first 1,000 are discarded as burn-in cycles. The estimation of one dataset composed of 60 items nested in 10 testlets and 500 examinees is completed within 20 minutes on a desktop with a 1.8 Ghz central processor unit, whereas IRT estimation usually takes no more than 10 minutes.

Following estimation, the point estimates and the estimated SEEs of the ability parameters from IRT and TRT models are compared. Parameter recovery from IRT and TRT model estimations is compared and evaluated in terms of bias (Equation 24), mean absolute error (Equation 25), root mean squared error (RMSE; Equation 26), mean theoretical SEE (Equation 27), and empirical SEE (Equation 28) of the ability estimates averaged across all replications. The equations for these statistics are shown as follows. The bias is presented as

$$\text{Bias}\left(\widehat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta\right), \tag{24}$$

with $\widehat{\theta}_r$ indicating the estimate of each repetition, R indicating the number of repetitions, and θ representing the true value of the variable. The mean absolute error is presented as

$$\text{mean abs error}\left(\widehat{\theta}\right) = \frac{1}{R}\sum_{r=1}^{R}\left|\widehat{\theta}_r - \theta\right|. \tag{25}$$

RMSE is presented as

$$\text{RMSE}\left(\widehat{\theta}\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \theta\right)^2}. \tag{26}$$

The mean theoretical SEE is

$$\text{SEE}\left(\widehat{\theta}\right) = I^{-1/2}\left(\widehat{\theta}\right), \tag{27}$$

with I indicating the information function of the estimate, and the empirical SEE is presented as

$$\text{SEE}_{\text{empirical}}\left(\widehat{\theta}\right) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \overline{\theta}_t\right)^2}, \tag{28}$$

with $\overline{\theta}_t = \frac{\sum_{t=1}^{R}\widehat{\theta}_t}{R}$. $\overline{\theta}_t$ represents the average of the best estimator $\widehat{\theta}_t$ over replications.

Information-Correction Procedure

The variance components are derived from the generalizability analysis based on Tables 1 – 3 that were described in the theoretical framework. The information-correction ratio (t_d) is calculated based on Equation 10. The square root of the inverse of this information-correction ratio ($t_d^{-1/2}$) is applied to adjust the SEE of ability parameter calibrated with the unidimensional IRT models (i.e., 1PL, 2PL, and 3PL IRT models). For each condition, these adjusted conditional SEEs are plotted against the ability scale and compared with SEEs from the IRT models as well as SEEs from the corresponding TRT models. The increase rates in SEE as a result of the information correction $\frac{\mathrm{SEE}_{\mathrm{IRT}} t_d^{-1/2} - \mathrm{SEE}_{\mathrm{IRT}}}{\mathrm{SEE}_{\mathrm{IRT}}}$ or as a result of the TRT modeling $\frac{\mathrm{SEE}_{\mathrm{TRT}} - \mathrm{SEE}_{\mathrm{IRT}}}{\mathrm{SEE}_{\mathrm{IRT}}}$ with the IRT SEE as the baseline are computed. SEID (Equation 18) is the dependent variable used to evaluate the effect of adjustment by the information-correction method. A SEID value close to zero indicates sufficient adjustment in error variance and hence a good performance of the information-correction method. To determine the effects of each manipulated factor on the performance of the information-correction method, descriptive and inferential statistics (analyses of variance, ANOVA) are presented to determine whether the observed differences across simulation conditions in the dependent variable are of statistical significance.

Methodology: A Real Data Example

Data

This large-scale test was administered to 827 examinees in Grade 3 to assess their reading skills. The test consists of 40 multiple-choice items nested in nine testlets. Each testlet is associated with a reading passage. Because one of the testlets contains only two items, which presents difficulty in producing an accurate estimate of the random testlet effect, those two items were deleted. A brief summary of the testlet structure is shown in Table 6.

Research Questions

1. What are the characteristics of the test in terms of LID and dimensionality?
2. Which type of response theory model fits the response data best?
3. How are the estimates from the IRT model compared with the estimates from the TRT model with the same number of item parameters by using Bayesian estimation through the MCMC procedure?
4. How are SEEs of proficiency estimates from the IRT model compared with SEEs adjusted by the information-correction method and SEEs from the TRT model?

Analysis: Conditional Independence Assessment

The local independence assumption of the IRT models is evaluated using Yen's (1984) Q_3 statistics, which is calculated from the correlation of the residuals of an item pair based on IRT models. For test forms that exhibit no or minor LID, the unidimensional IRT models are more parsimonious and might produce more accurate estimates than their TRT counterparts. Thus, it is necessary to know whether the test design and the item format conform to the characteristics of testlets and would allow the applications of testlet models and the information-correction method. The distributional characteristics of the Q_3 statistics of each testlet are computed and compared with the expected value of Q_3 statistics. To understand which simulations are closest to the real data, the Q_3 statistics of one dataset in each condition in the 3PL context are also estimated.

Table 6 The Structure of Reading Comprehension Test

Testlet	1	2	3	4	5	6	7	8
Number of items	6	3	3	6	6	5	3	6

Analysis: Factor Analysis

To evaluate the dimensionality, the exploratory factor analysis is conducted on a tetrachoric correlation matrix of response variables on the test responses. For the testlet model, the general factor systematically affects examinees' performance in the tests, and the factors other than the primary factor can be regarded as secondary factors limited within the testlet level.

The exploratory item analysis is implemented in TESTFACT 4.0 (Wilson, Wood, & Gibbons, 1991), by using all the information in the data matrix through MMLE. First the smoothed tetrachoric correlation matrix is obtained, and then TESTFACT performs a principal factor analysis on the correlation matrix using the minimum squared residuals method. Varimax rotation is chosen in this example. As suggested by many researchers (e.g., Gorsuch, 1983), examination of scree plots is used for determining the number of factors.

Analysis: Model Selection

Although the testlet models yield more accurate results when fitting the data with certain magnitude of testlet effect, they are overfitting to the data with minimum amount of LID. Likewise, the 2PL models are parsimonious and have a better model fit compared with the 3PL model if the pseudoguessing parameter values are not significantly different. The four types of models (i.e., the 2PL IRT, the 2PL TRT, the 3PL IRT, and the 3PL TRT) are potential options to estimate this test dataset. These four models are expected to lead to different solutions and interpretations, which need to be evaluated on the basis of model fits.

In this study, the Deviance Information Criterion (DIC) is used to select the model with the best fit. DIC is a built-in function in WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003a), in which parameter estimation is implemented. Compared with other model fit indices of AIC or BIC, DIC is effective in complex hierarchical models where parameters may outnumber observations (Gelfand & Dey, 1994). DIC defines not only a measure of fit but also a measure of complexity (Spiegelhalter, Best, Carlin, & van der Linden, 2002). The model with the minimum DIC is the one preferred. At the same time, Spiegelhalter, Thomas, Best, and Lunn (2003b) also suggested, as a rule of thumb, that a difference of at least 3 – 7 could be considered significant (p. 613).

Analysis: Parameter Estimation

Model estimation was implemented in WinBUGS through the MCMC procedure. As suggested by the results of the previous simulation studies, the Bayesian procedure is relatively robust to different specifications of prior distributions so long as the parameters are well identified and not too extreme (Gifford & Swaminathan, 1990; Swaminathan & Gifford, 1982, 1985, 1986). In that case, the prior distributions are specified based on the convention of a typical achievement test (Bradlow et al., 1999; DeMars, 2006; Li, Bolt, & Fu, 2005; Wainer et al., 2000), and were given by

$$\begin{aligned}
\theta_i &\sim N(0,1)\\
\alpha_j &\sim \text{logN}(0,0.25)\\
\beta_j &\sim N(0,4)\\
\omega_j &\sim beta(5,17)\\
\gamma_{id(j)} &\sim N\left(0,\sigma^2_{\gamma d(j)}\right),
\end{aligned} \tag{29}$$

where θ_i is the person proficiency of person I, α_j is the item discrimination on item j, β_j is the item difficulty of item j, and ω_j is the pseudoguessing parameter of item j. The hyperprior distribution of the variance of testlet effect $\sigma^2_{\gamma_{d(j)}}$ is assumed to be an inverse gamma distribution with α and β parameters both set at 1, $\sigma^2_{\gamma_{d(j)}} \sim \Gamma^{-1}(1,1)$. Two chains of item difficulty parameters with very divergent starting values ($-2, -2, \ldots \ldots, -2$) and ($2, 2, \ldots \ldots, 2$) are run for each model, and the program is requested to generate the other starting values. The purpose for running two chains is to ensure the convergence of two chains when the estimates reach stationary.

Among several convergence diagnostic criteria that are available in WinBUGS, dynamic trace lines, history, and Gelman – Rubin convergence statistics are often used for the purpose of convergence diagnosis. The diagnostic graphs and statistics indicate that the two chains achieve convergence within the first 2,000 iterations. To be conservative, the

first 4,000 burn-in cycles are discarded. When the 3PL IRT and the 2PL and 3PL TRT are fit onto the data, the adaptive box is automatically checked, which suggests that WinBUGS is using a complex sampler such as a Metropolis sampler. In this circumstance, the default number of burn-in iterations is 4,000. For the 2PL TRT model, the first 4,000 cycles are burned in. For the 3PL IRT and TRT models, the diagnostic indictors seem to suggest that convergence does not happen until the 5,000th iteration. In this case, the first 6,000 iterations are discarded as burned-in cycles. The estimated standard errors and the point estimates of ability parameter are extracted from the statistics of the posterior distributions.

Analysis: Information Correction

The information-correction ratios are estimated by following the steps shown in Tables 1 – 3. The standard errors of the proficiency estimates from the selected IRT model are corrected by the estimated information-correction ratios. The effect of the correction is evaluated by comparing the IRT SEEs, the IRT SEEs adjusted by the correction ratios, and the TRT SEEs. To understand the possible differences and similarities between the real test analysis and the simulated data analysis, comparisons are made in terms of LID (indicated by Q_3), the testlet length, the balance of testlet length, and the discrepancy between the adjusted IRT SEE and TRT SEE (indicated by SEID).

Results: Simulation Study

Ability Parameter Recovery

It is shown that in the case of independent item test ($\sigma_\gamma^2 = 0$) the average mean absolute errors from the TRT ability estimates are generally higher than those from the IRT ability estimates, which indicates the overparameterization of TRT models. However, for the responses simulated with significant LID, the mean absolute errors from the TRT ability estimates are lower than those from the IRT estimation, which indicates that TRT might have a better model fit than IRT in this situation. It seems that the increase in LID will lead to the increase in the mean absolute errors for both IRT and TRT models. Long testlets seem to have higher mean absolute errors than short testlets, especially in the conditions where the magnitude of LID is large, which is aligned with the results from the previous studies that short testlets tend to have moderate effects on estimation even though the independent item models are misspecified to the testlet dataset. RMSE demonstrates a similar pattern as in the mean absolute error: Namely, IRT provides more accurate estimates for tests with no LID, but TRT offers more accurate estimates for tests with LID; accuracy decreases with the increase in LID, and long-testlet tests tend to result in less accurate estimates than short-testlet tests. The mean theoretical SEEs from the IRT ability estimates are similar across conditions with the same number of item parameters, because IRT models do not account for LID among items within the testlet. The mean theoretical SEEs from TRT ability estimates increase as LID goes up. In addition, the short-testlet test seems to result in lower mean theoretical SEE than the long-testlet test. These observations suggest the overestimation of precision in the cases of high LID or in the long-testlet cases given that other factors are equal. Mean theoretical SEE is only slightly higher for unbalanced tests than balanced tests in terms of testlet length, perhaps because the parameters are less easy to estimate when there are very few items in a testlet.

Information Correction

The conditional SEE from IRT models (SEE_{IRT}) and the conditional SEE adjusted by the information-correction terms ($SEE_{IRT}t^{-1/2}$) are compared with the conditional SEE from TRT models (SEE_{TRT}). Figures 1 through 9 illustrate the results of the comparison. The discrepancy between SEE_{IRT} and SEE_{TRT} becomes larger when LID increases or the testlet length decreases. However, SEE_{IRT} can always be adjusted to the value that is close to SEE_{TRT} by using the information-correction ratio, which suggests that the information-correction method is effective for this purpose. The conditional standard error plots further show that when LID is zero, $SEE_{IRT}t^{-1/2}$ is lower than the targeted SEE_{TRT} across the ability scale, but when LID is substantial, $SEE_{IRT}t^{-1/2}$ is very close to SEE_{TRT} conditional on θ values in the middle part of the scale. $SEE_{IRT}t^{-1/2}$ conditional on extreme values on the θ scale tends to be higher than SEE_{TRT}, which suggests that the information correction presents satisfactory performance for the conditions with substantial testlet effects, but overcorrection may occur to SEE conditional on extreme ability values. The unbalanced conditions seem to result in better correction than the balanced conditions. 3PL models have better information-correction performance than 2PL models, which in turn result in better correction than 1PL models.

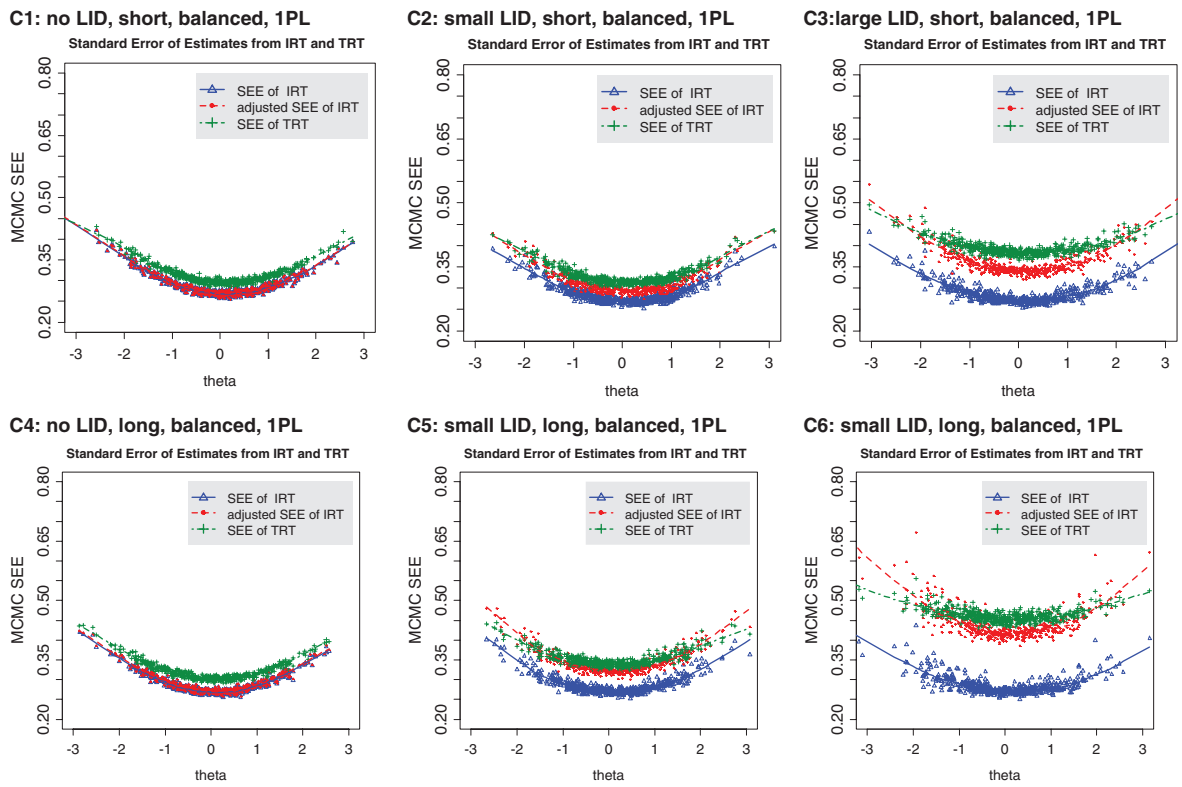

Figure 1 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for balanced one-parameter logistic.

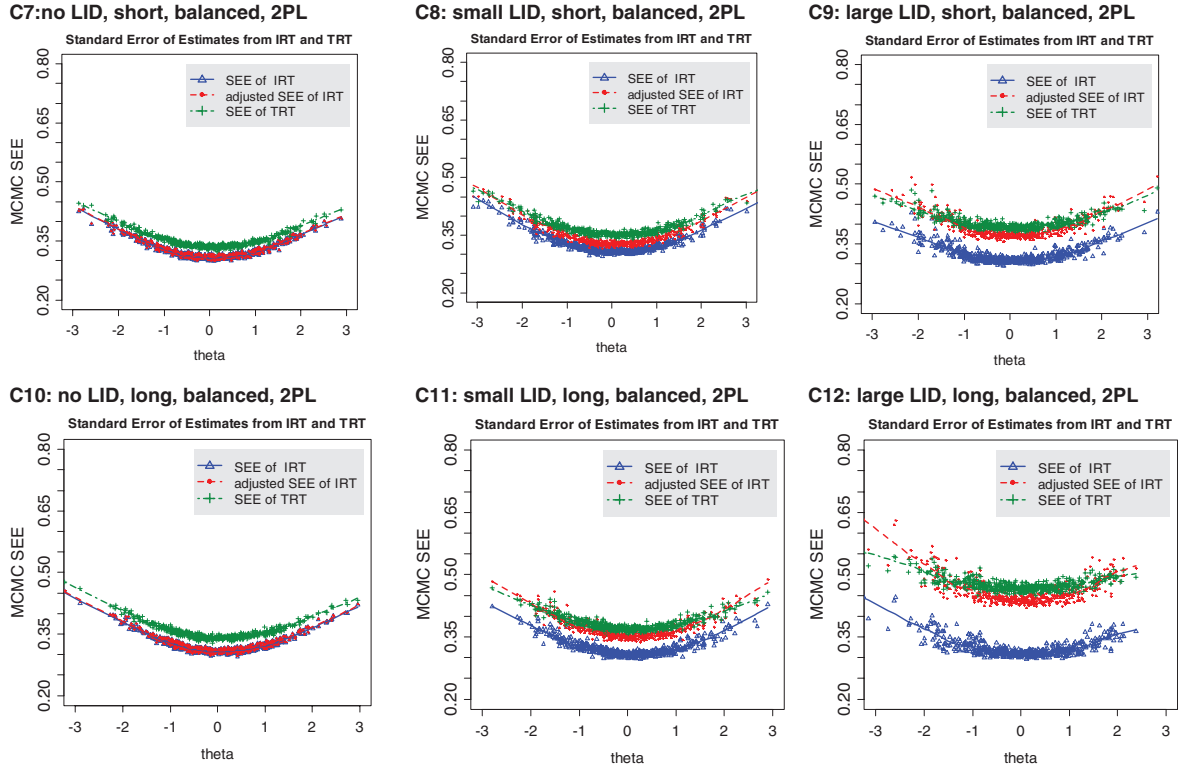

Figure 2 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for balanced two-parameter logistic.

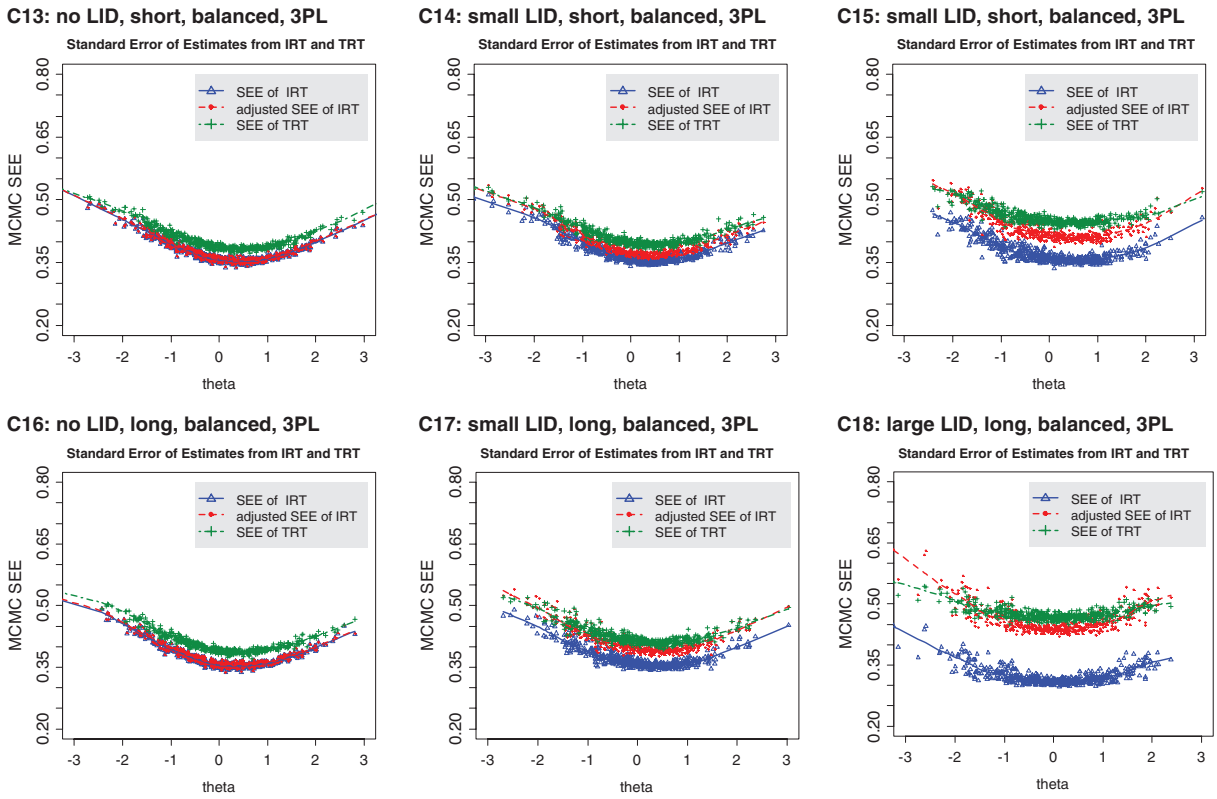

Figure 3 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for balanced three-parameter logistic.

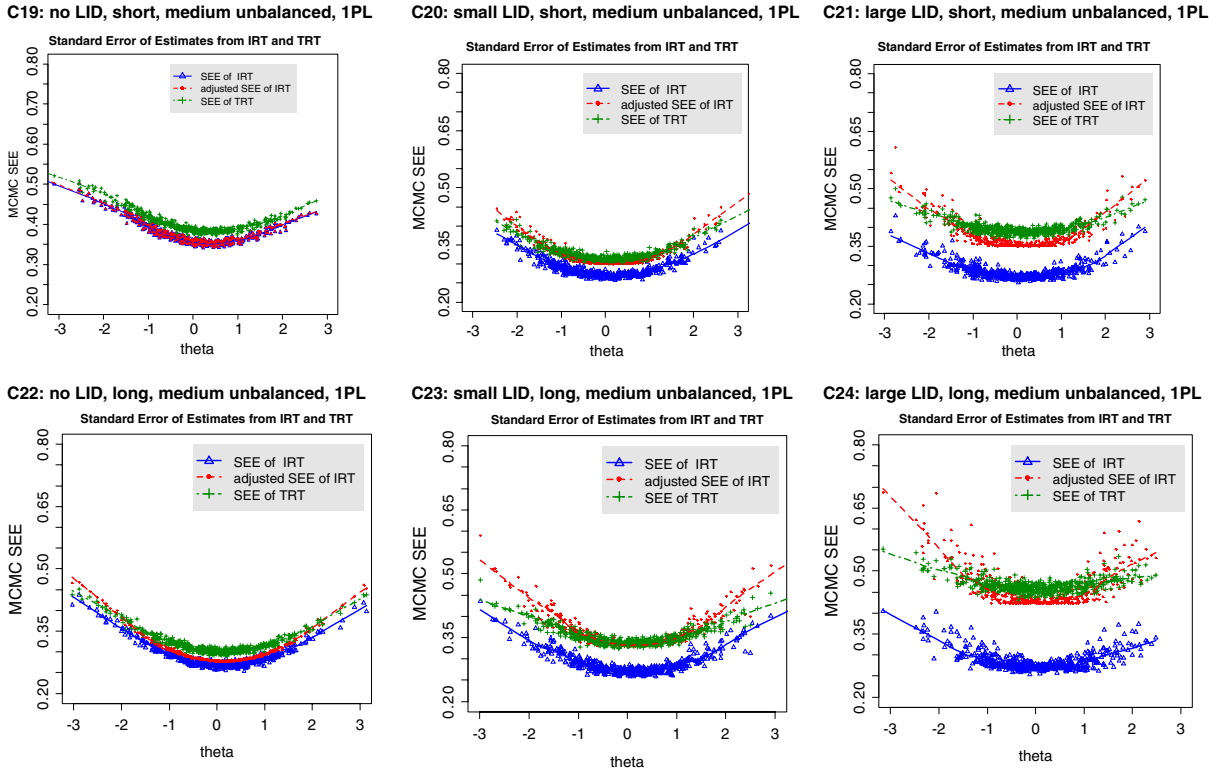

Figure 4 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for medium unbalanced one-parameter logistic.

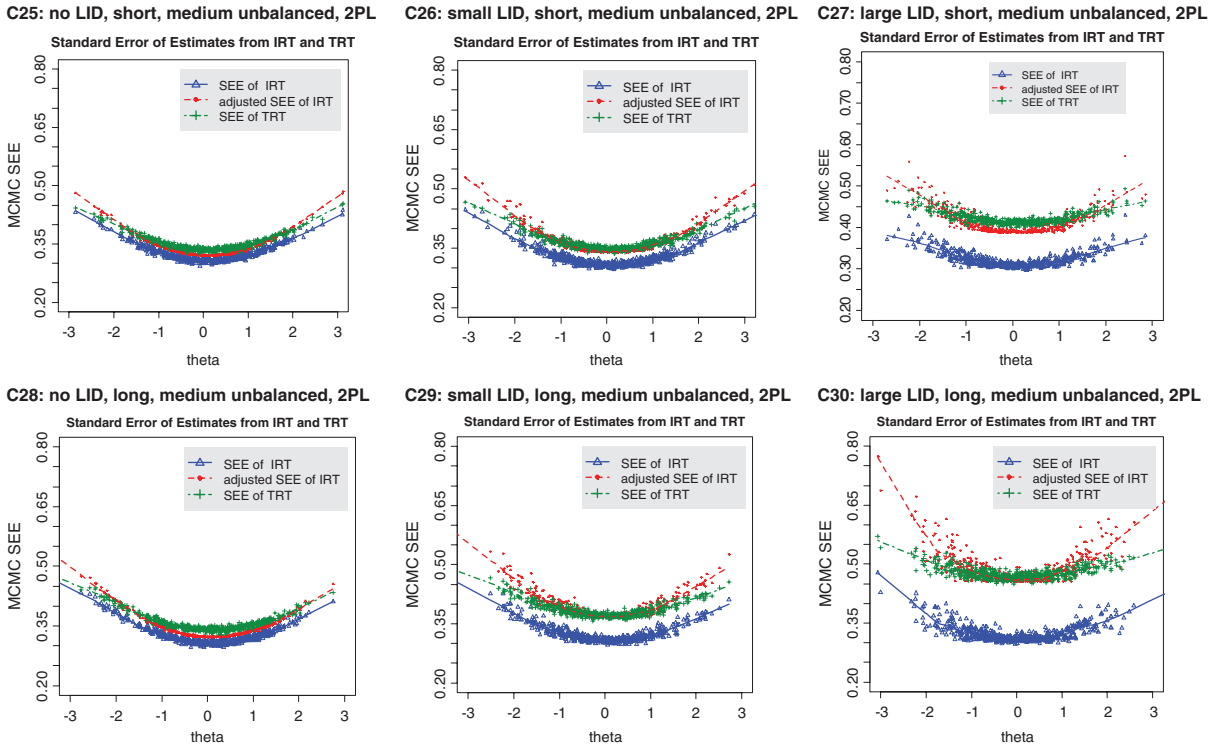

Figure 5 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for medium unbalanced two-parameter logistic.

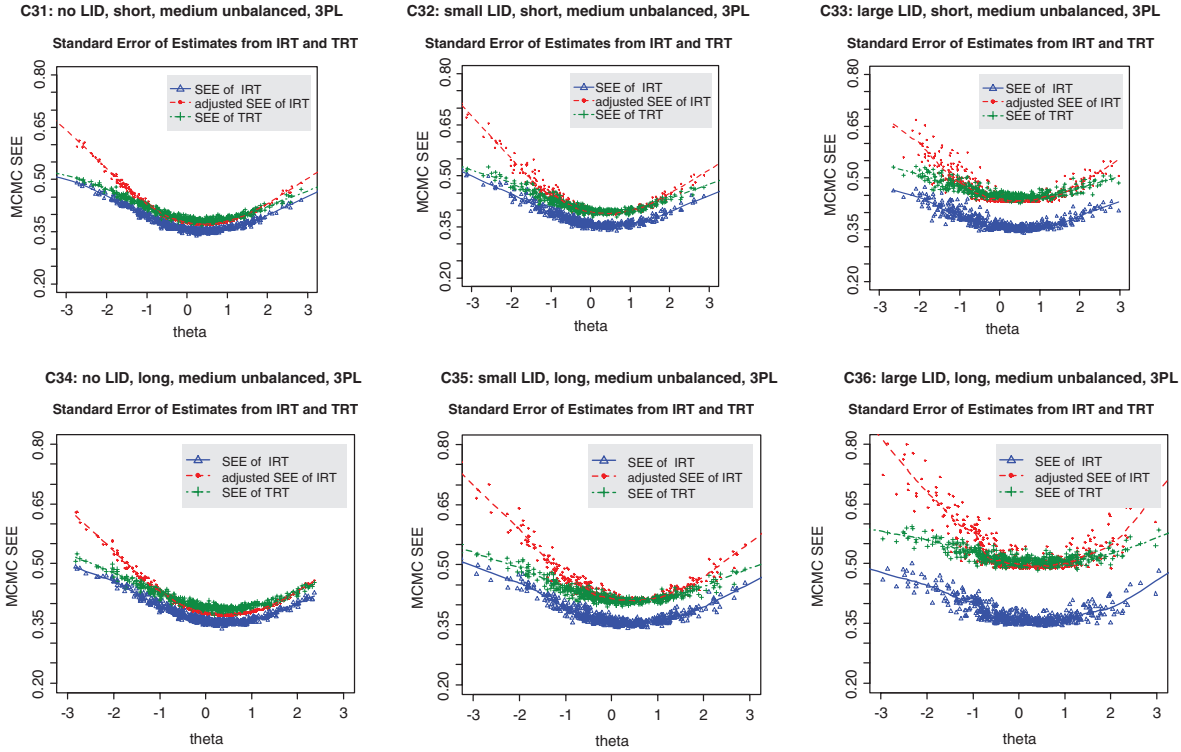

Figure 6 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for medium unbalanced two-parameter logistic.

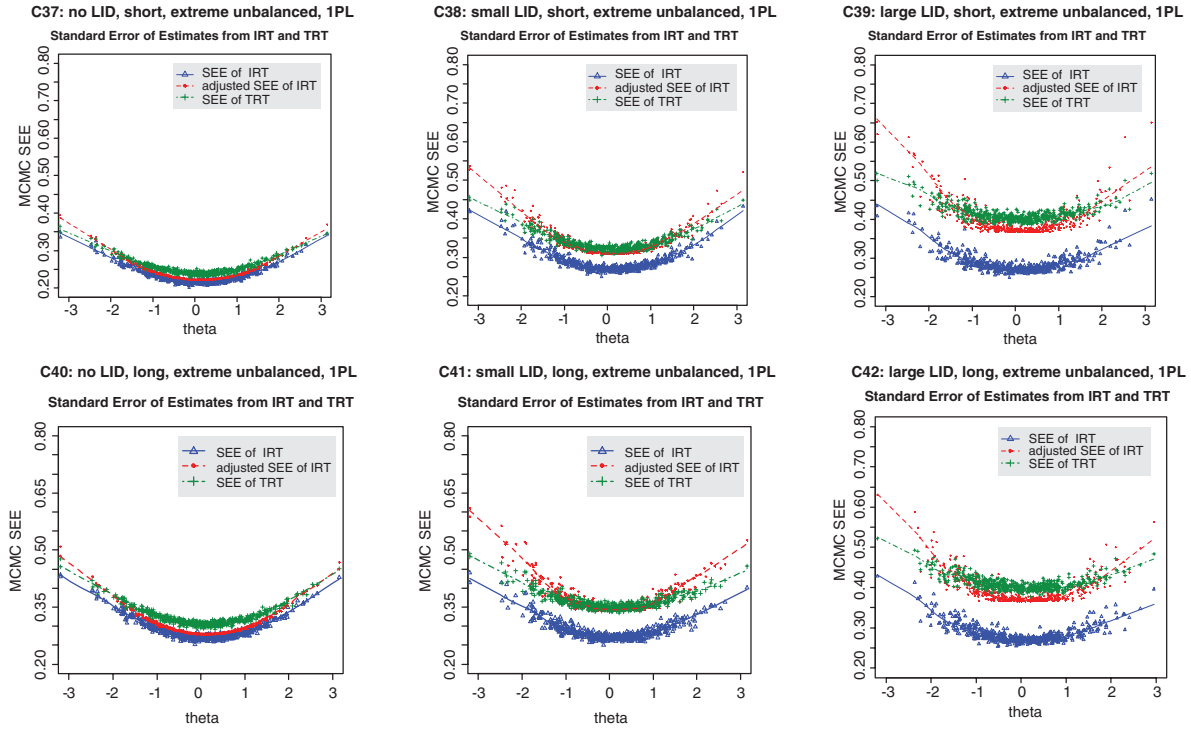

Figure 7 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for extreme unbalanced one-parameter logistic.

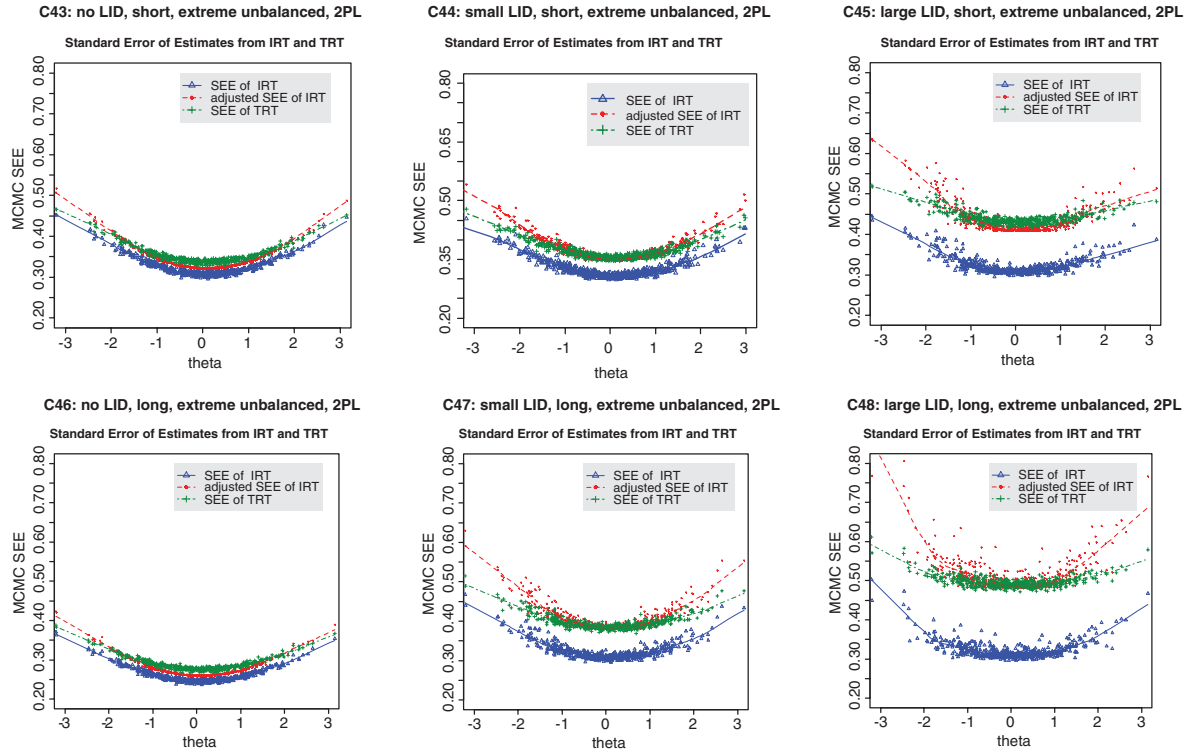

Figure 8 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for extreme unbalanced two-parameter logistic.

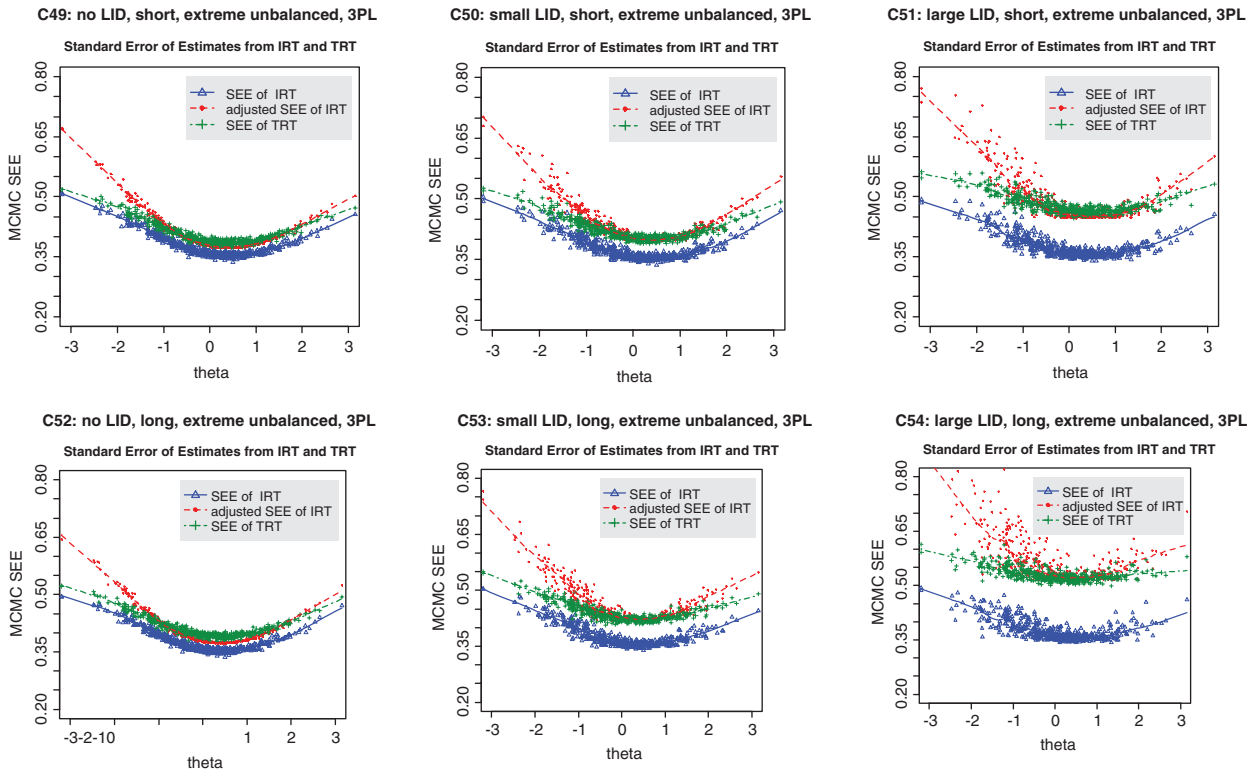

Figure 9 Standard error of estimate from item response theory before and after adjustment, and standard error of estimate from testlet response theory for extreme unbalanced three-parameter logistic.

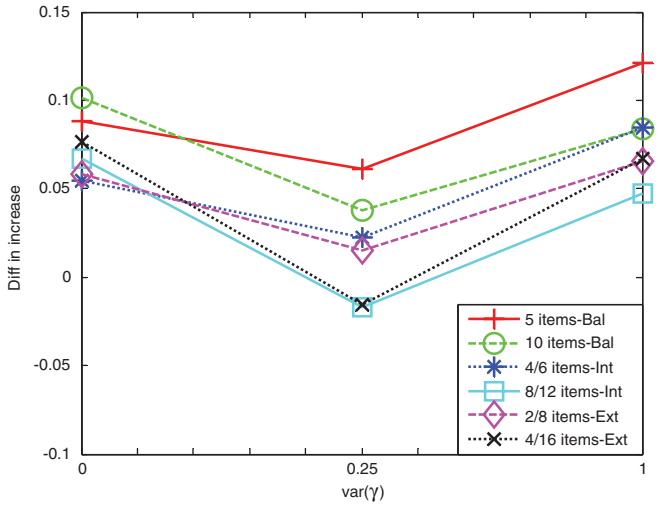

Figure 10 Standard error increase discrepancy from one-parameter logistic.

The mean SEID (i.e., the SEID statistics averaged across all examinees) represents the effect of information correction in general for a particular condition. However, undercorrection for some SEEs and overcorrection for other SEEs may cancel each other out and result in a low mean SEID value, as if all SEEs were appropriately corrected. According to the mean SEID plots in Figures 10 through 12, the mean SEID appears to be close to zero when LID is moderate ($\sigma_\gamma^2 = 0.25$) but comparatively deviated from zero when LID is zero ($\sigma_\gamma^2 = 0$) or large ($\sigma_\gamma^2 = 1$). This implies that the information-correction method might perform best in the conditions with moderate LID. However, the conditional SEE plots do not show better adjustment for moderate LID conditions than their large LID counterparts. Mean SEID values that are more

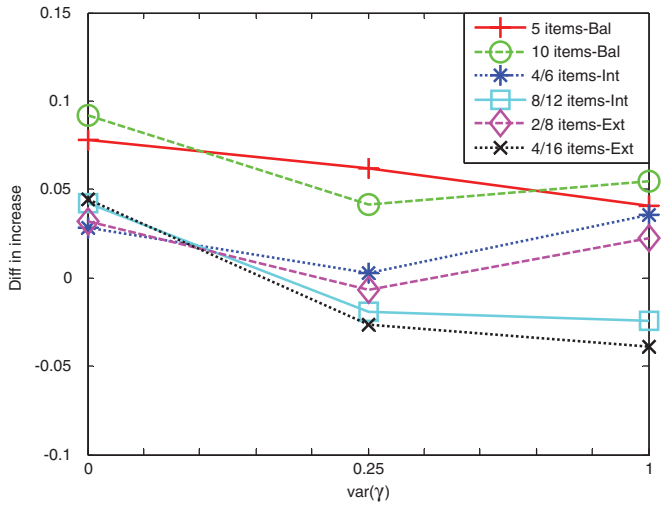

Figure 11 Standard error increase discrepancy from two-parameter logistic.

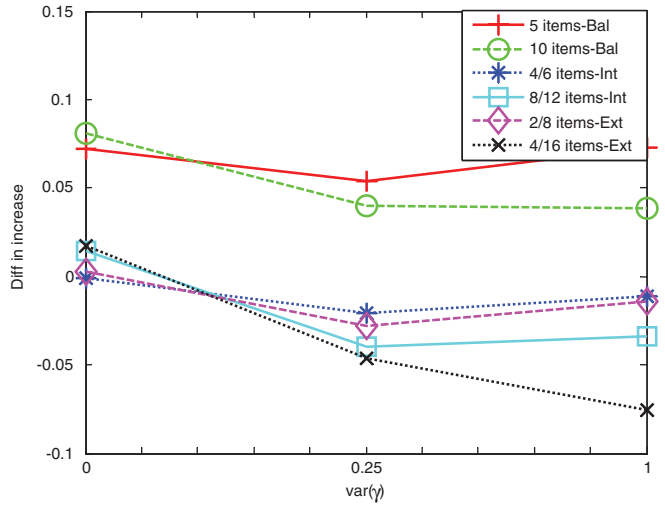

Figure 12 Standard error increase discrepancy from three-parameter logistic. *Note*. The standard error increase discrepancy: the difference between the standard error of estimates of proficiency from item response theory models adjusted by information-correction ratio compared against those from testlet response theory models. $\frac{\mathrm{SEE}_{\mathrm{IRT}} t_d^{-1/2} - \mathrm{SEE}_{\mathrm{TRT}}}{\mathrm{SEE}_{\mathrm{IRT}}}$.

Table 7 Means, Standard Deviations, and *T*-scores of Q_3 Statistics within Testlets

Testlet	1	2	3	4	5	6	7	8
Mean	.0252	.0824	.0852	.0023	.0063	.0304	.0401	.1162
SD	.0508	.0491	.0903	.0501	.0395	.0357	.1120	.0440
T-score	1.0273	2.2260	1.2438	.5847	.8426	1.6078	.5998	3.2549

Note. The expected value of Q3 is $-1/(38-1) = -.0270$.

deviated from zero on conditions with large LID may be attributed to overcorrection of $\mathrm{SEE}_{\mathrm{IRT}}$ on extreme θ values. With only a few exceptions, the mean SEID statistic also suggests that the adjustment effect improves as the degree of the unbalance of the testlet length increases, which is consistent with what has been observed from the conditional SEE plots. Based on both mean SEID and conditional SEE plots, the long-testlet test tends to result in a better overall adjustment than the short-testlet test.

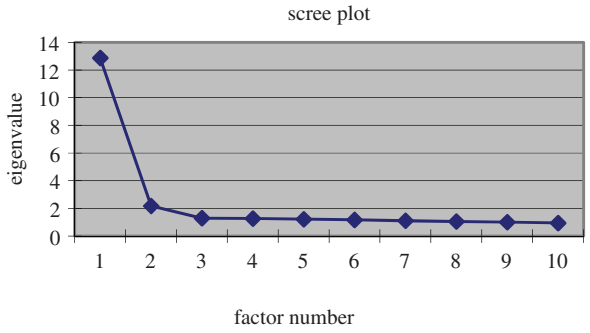

Figure 13 Scree plot from principal component analysis on tetrachoric correlation matrix.

The ANOVA results indicate that the three factors being manipulated (i.e., LID, testlet length, and the balance of the testlet length) and their interactions in this study account for more than 99% of the total variance in the dependent variable. Based on the *p*-values of the *F* tests, all of the three factors and their interactions are statistically significant, which implies significant effects on the adjustment of random errors using the information-correction method. Balance of the testlet length, LID, and the interaction between these two terms explain a large proportion of the total variance.

Results: Real Data Analysis

Conditional Independence

Although Q_3 is a correlation between residuals of an item pair, Q_3 has a tendency to be slightly negative when the CI holds (Chen & Thissen, 1997; Yen, 1984, 1993). Yen (1993) demonstrated that the expected value of Q_3 is approximately $-1/(n-1)$, and n denotes the number of test items. The expected value for Q_3 can be used as a criterion for comparing the overall level of local dependence of within-testlet item pairs. When CI holds, the average of Q_3 from within-testlet item pairs will be similar to the expected values of the Q_3.

Table 7 shows that average within-testlet Q_3 statistics have more positive values compared to the expected values of Q_3. This suggests that CI is violated. By referring to the *t*-scores of the observed Q_3 statistics, the magnitude of the differences between the observed values and the expected value of Q_3 seem to be approximately one SD or even larger, except for testlets 4, 5, and 7, where these differences are moderate.

To understand which simulation condition is closest to the response matrix of the real test, the Q_3 statistics of one simulated dataset in each condition in the 3PL context are estimated. By comparing both means and *t*-scores of Q_3 statistics within testlets, it is found that the Q_3 pattern of the real test data is somewhere between Conditions 32 and 33. Conditions 32 and 33 share the features of short-testlet length and intermediate level of unbalance in testlet length. The only difference between them is that in Condition 32, the dataset is generated with a small variance of testlet effect, whereas in Condition 33, the dataset is generated with a large variance of testlet effect.

Dimensionality

Through the exploratory factor analysis, the first component accounts for over 40% of the overall variance in the dependent variable. The largest eigenvalue is about six times as large as the second largest. According to the scree plot presented in Figure 13, the eigenvalue of the first component is significantly larger than those of all the other components. All these suggest that one factor is dominant in this dataset. From the table of factor loadings, almost all items load highly on the first factor compared with loadings on other factors, which also confirms that this test has a dominant dimension (see Table 8).

Model Selection

A minimum value of DIC indicates a parsimonious model with good model fit. As a result of MCMC estimation, it turns out that the 3PL TRT model is preferred, as it has the smallest DIC among the four models (Table 9). Therefore, the 3PL

Table 8 Eigenvalues of the Principal Components

Component	Eigenvalues	% of variance	Cumulative %
1	12.873	40.267	40.267
2	2.180	6.111	46.378
3	1.292	3.298	49.676
4	1.268	2.864	52.540
5	1.224	1.929	54.469
6	1.175	1.359	55.828
7	1.108	1.086	56.914
8	1.049	0.671	57.585
9	1.008	0.471	58.055
10	0.944	0.448	58.504

Table 9 Deviance Information Criterion Values for Four Models

Model	IRT	TRT
2PL	33640.5	32960.5
3PL	33514.9	32921.0

IRT model is selected to calibrate the parameters. The estimated standard errors of ability estimates are adjusted by the ratio of variances from the response matrix and compared with the standard errors of ability estimates calibrated through the 3PL TRT model.

Parameter Estimation

After the 3PL IRT and 3PL TRT models are selected for ability estimation, the next step is to check convergence in MCMC estimation. It is done by examining whether the simulated Markov chain converges to a stationary distribution. A random subset of parameters is selected for this purpose. A large number of iterations are usually required to ensure the convergence and stable estimates for the complicated models such as 3PL IRT and TRT models (Sinharay, 2003). Three approaches are generally used in assessing convergence. The first approach is to examine the history plot, which shows the full history of the sample values for the parameter being monitored. Second, we can look at trace plots of the sample values versus iteration to see when the simulation appears to have stabilized. If the chains starting from divergent initial values in the trace plot or history plot appear to be overlapping one another, we have evidence to believe that convergence has taken place. The third diagnostic approach is the Gelman – Rubin index. For the Gelman – Rubin plots, the width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio (pooled/within) is red (Brooks & Gelman, 1998). The convergence is indicated when the blue and green curves overlap and the red curve hovers around 1. Based on the history plots, trace plots, and the Gelman – Rubin plots, the convergence is achieved after 6,000 iterations for 3PL IRT estimation and 10,000 iterations for 3PL TRT estimation. For the 3PL IRT model, 50,000 iterations are run in the numerical implementation. The first 6,000 iterations are discarded as burn-in cycles, so the parameters are estimated from the posterior distributions based on the 6,001st to the 50,000th iteration. For 3PL TRT model estimation, the posterior distributions are estimated based on the 10,001st to the 70,000th iterations. The means of the Bayesian posterior distributions are used for item parameters in the calculation.

The correlation statistics show that the two sets of item parameter estimates are highly correlated. Item difficulty (b) estimates obtained through IRT are almost perfectly correlated with those from TRT ($r = .995$). Item discrimination (a) estimates are highly correlated ($r = .882$). Agreements on person proficiencies (θ) are also high ($r = .994$). In contrast, there is less agreement for guessing estimates ($r = .756$). The differences between the estimates from the IRT and TRT models indicate their impact on parameter estimation.

Table 10 represents the variances of the testlet effect variable in each testlet. The magnitudes of the testlet effect range from small to moderate. For example, testlets 1, 4, and 5 have small variances of testlet effect variable. There were substantial effects for testlets 2, 7, and 8. All estimates of the variances of testlet effects have acceptable standard errors. By

Table 10 Means and Standard Deviations of Posterior Distributions of Testlet Effect Variances and Correction Ratios

Testlet	1	2	3	4	5	6	7	8
Number of items	6	3	3	6	6	5	3	6
$\sigma^2_{\gamma d(j)}$	.25	.71	.48	.19	.19	.38	.65	.61
SE	.06	.19	.12	.05	.04	.09	.19	.11
Mean of Q_3	.0252	.0824	.0852	.0023	.0063	.0304	.0401	.1162
Ratio*	1.1633	1.0683	1.0683	1.1633	1.1633	1.1325	1.0683	1.1633

Note. *Ratio is the G-theory correction ratio.

Table 11 Mean Squares, Variance of the Random Variance Components From G-theory Analysis

Source of variation	df	MS	σ^2
Examinees	826	1.7395	0.0396
Testlets × examinees	6616	0.2282	0.0128
(Items: testlets) × examinees	24810	0.1682	0.1682

Note. $r = \sum_d \frac{k_d^2}{n} = 5.1579$; $t = \frac{n-r}{m-1} = 4.6917$.

comparing the estimates of σ^2_γ with the mean of Q_3, we may notice that high σ^2_γ statistics are associated with high Q_3, whereas low σ^2_γ tends to be associated with low Q_3.

Information Correction

Based on the analysis above, this test is characterized by short testlets and LID that ranges from moderate to large. Testlet lengths are unbalanced to an intermediate extent across the test. 3PL IRT and TRT models are used for ability estimation. Both the test characteristics and the Q_3 pattern are close to those of simulation Conditions 32 or 33. By referring to Figure 12 for the condition with moderate or large LID ($\alpha^2_\gamma = .25$ or $\alpha^2_\gamma = 1$) and four or six items in each testlet, the mean SEID statistic values are close to zero. It implies that on average, IRT SEE can be adjusted to be very close to TRT SEE. By referring to Conditions 32 and 33 of Figure 6, the adjusted IRT SEE and TRT SEE are almost overlapping when θ values range from −1.5 to 1.5 on the ability scale, whereas IRT SEE seems to have been overadjusted as compared against the TRT SEE for extreme θ values on the ability scale. Thus, it is speculated that IRT SEE could be corrected to a level that is close to TRT SEE in the test of this example.

The partition of variances is shown in Table 11. The correction ratios are listed in Table 10. The correction ratios that are specific to each testlet are calculated as a function of the testlet length. The conditional standard errors of proficiency estimates from 3PL IRT, 3PL TRT, as well as those from 3PL IRT adjusted by the correction ratios, are plotted in Figure 14.

It is noticed that IRT SEEs have been increased as a result of adjustment. The mean SEID is −.0459, which indicates that on the average, the adjusted IRT SEE is 5% higher than the targeted TRT SEE based on IRT SEE. It suggests a satisfactory adjustment effect in general compared with the mean SEID values in the simulation study. However, by referring to the conditional SEE plots (Figure 14), the IRT SEEs conditional on the TRT ability estimates between −1 and 1 seem to be underadjusted compared with TRT SEE, but the magnitude of this underadjustment is very small and no more than .05. In contrast, the IRT SEEs conditional on TRT ability estimates beyond either −2 or 2 seem to be overadjusted, but the two extreme ends of the ability scale include less than 2% of the examinees in this test. By looking at the SEID statistics conditional on the ability estimates (Figure 15), the majority of the SEID values range between 0 and .3 for ability estimates between −1.5 and 1.5.

In Condition 33 of the simulation study, the IRT SEEs conditional on θ values between −1 and 1 have been adjusted to be overlapping to TRT SEEs, whereas in this real data example, the IRT SEEs conditional on this part of the ability scale are a little bit lower than the TRT SEEs. One possible explanation for this difference is that all testlets in the simulation study are generated to have the same magnitude of the variance of testlet effect, but LID in this real example analysis is unequal across testlets. It is possible that the correction ratio is not only a function of the testlet length but also of the error variances specific to each testlet.

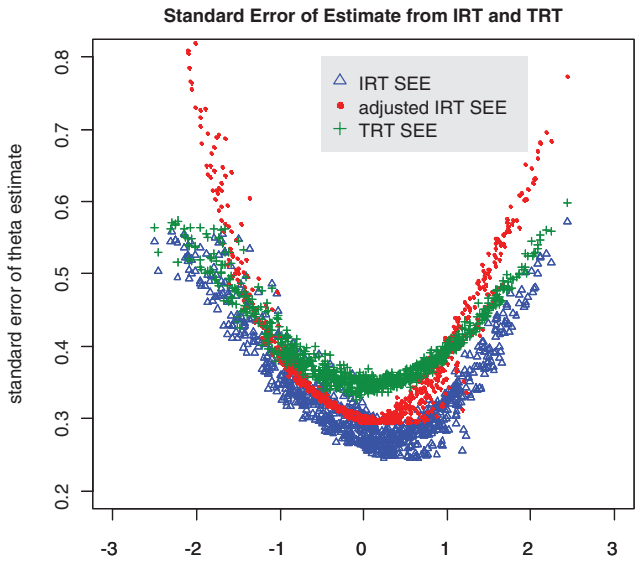

Figure 14 Standard errors of ability estimates.

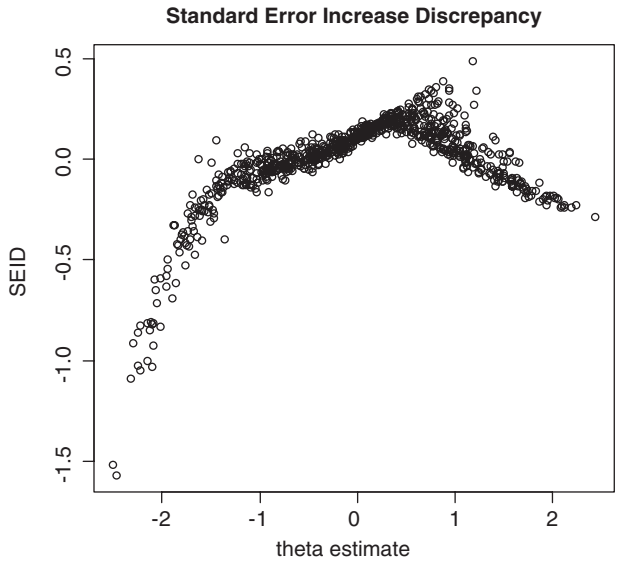

Figure 15 Standard error increase discrepancy conditional on testlet response theory ability estimates.

Conclusions and Discussion

The work conducted in this research extends the information correction of multiple ratings proposed by Bock et al. (2002) and demonstrates how GT and IRT could be implemented sequentially to obtain more accurate precision estimates in a testlet-based test. Through the simulation study, the performance of the information-correction method is examined in the 1PL, 2PL, and 3PL IRT contexts with the varying magnitude of LID, testlet length, and balance of testlet length. Through a real data analysis, the measurement errors yielded by the information-correction method are compared to those from the TRT model.

Implications for Testing Practices

This research addresses the problem that the conditional-independent IRT models that do not account for LID in testlets would lead to the underestimation of the measurement errors of proficiency parameters. This issue can be critical for

scoring in high-stakes tests or for proficiency classification. In complex computerized simulations, the precision estimates can affect the IRT evidence accumulation process. Therefore, it is necessary to have a relatively accurate estimate of the measurement precision and to quantify the local dependency of testlets. Although some testlet models have demonstrated satisfactory performance in terms of model-data fit and parameter recovery, each of them has limitations. The GT model has not been sufficiently developed to connect continuous latent values with discrete scores. Models of the IRT approach are complex and usually take a long time to converge because of the ways they are currently estimated.

In this study, it is shown that the information-correction method is efficient and straightforward, as it is easy to derive the error variances of person parameters in either the testlet design or the independent item design from the GT analysis as well as the precision estimates from IRT models. Given the corresponding relationship in error variance ratios between the generalizability models and response theory models, it should be reasonable to apply the information-correction term to testing practices.

The simulation study provides evidence that the underestimated measurement errors from IRT estimation could be adjusted to the appropriate level through the information-correction method despite the varying LID, testlet length, balance of testlet length, and number of the item parameters in the model. The expected values of error variances from the TRT estimation can be assumed as the benchmark because TRT models account for LID and thus can produce more accurate estimates about the testlet datasets. Given the robustness of variance ratios, estimation of the information correction should be adequate for practical work.

In addition to demonstrating the adequate performance of the information-correction method, results from ANOVA show the impact from the three factors (i.e., LID, testlet length, and the balance of the testlet length) on information correction. All three factors and their interactions present as statistically significant and account for more than 99% of the total variance in the dependent variable. In other words, the information-correction method is more effective for tests with certain characteristics. The information-correction method seems to perform better on long-testlet tests than short-testlet tests, better on tests with LID than the tests without LID, and better on tests with unbalanced testlet lengths than tests with balanced testlet length. The 3PL model context seems to have more satisfactory adjustment results than the 2PL context, which in turn has better adjustment effect than the 1PL context. Although the results of the significance tests rely upon the dependent variable that has been selected, this analysis has roughly depicted a picture about how the information-correction method performs in each situation.

The real data example has provided more details about the information-correction procedure. By comparing the real test and the simulated tests, it is shown how closely the error variance from a real data example can be adjusted to the results we would expect. In addition, it allows an investigation into how the correction coefficients work on the measurement error of each examinee's ability parameter when the calibration of item parameters is involved. It is noteworthy that diagnosis tests are necessary to detect LID and dimensionality so as to ensure that the correction procedure is applicable.

Limitations and Directions for Future Research

Presented in this study are the initial studies on the information-correction methods, so it is beset with certain limitations. First, as shown in both the simulation study and the real data analysis, the standard errors of proficiency estimates given extreme proficiency values have been overcorrected in many conditions. Because the correction ratio term is a function of the testlet length and is applied to all examinees across the ability scale, the standard errors that are already highly conditional on extreme ability values will be magnified with the multiplicative coefficient. In future studies, we may cut the ability scale into intervals and estimate correction ratio for each interval. Alternatively, we may also build a correction term as a function of not only the testlet length but also ability values.

In order to focus on the change in SEE, the estimation procedure was simplified by fixing all item parameter values, but in the hierarchical Bayesian framework, the estimation of the item parameters will affect the estimation of ability parameters. Therefore, we may also request the program to estimate the item parameters in the follow-up simulation studies. Another limitation of the simulation study is that in each condition, the correct model was used to estimate ability parameters. For example, the 3PL TRT model was used to estimate the dataset generated with 3PL TRT. A future topic for research can be the consequence of fitting misspecified TRT models to data.

The response datasets were generated with equal LID for each testlet in the simulation study, but in the real test, LID often varies across the testlets. It seems that the correction results from the real data analysis are somewhat different from

what was expected based on the simulation study. Thus, we may consider the simulation conditions of unequal LID in the future.

The expected value of TRT random error was treated as the benchmark in this study to evaluate the adjusted standard error, because TRT models provide better parameter recovery than IRT models based on the simulation studies. However, it is also true that on some occasions when LID is zero in the response matrix, TRT models provide less accurate estimates than IRT, so this is the limitation when a TRT model is used as the true model. It is worth further investigation for a research design in which effects from the confounded variables could be cleared.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement 21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement*, *26*(4), 364–375. https://doi.org/10.1177/014662102237794

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153–168. https://doi.org/10.1007/BF02294533

Brennan, R. L. (1992). *Elements of generalizability theory* (revised edition). Iowa City, IA: ACT.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Researcher*, *16*, 14–20. https://doi.org/10.1111/j.1745-3992.1997.tb00604.x

Brennan, R. L. (2001). *Generalizability theory*. New York, NY: Springer.

Briggs, D., & Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155. https://doi.org/10.1111/j.1745-3984.2007.00031.x

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

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

Cornfield, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*, *41*, 654–661. https://doi.org/10.2105/AJPH.41.6.647

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, *57*, 373–399. https://doi.org/10.1177/0013164497057003001

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145–168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

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

Gelfand, A. E., & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(3), 501–514.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436. https://doi.org/10.1007/BF02295430

Gifford, J.A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement*, *14*(1), 33–43. https://doi.org/10.1177/014662169001400104

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (Research Report No. RR-13-32). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory*. Boston, MA: Kluwer Nijhoff.

Jiao, H., & Wang, S. (2008). *Comparison of estimation methods of one-parameter testlet models.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, *6*(3), 311–321.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont Portfolio Assessment Program: Findings and implications. *Educational Measurement: Issues and Practice*, *13*(3), 5–16.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education*, *12*(3), 237–255.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*, *29*(5), 340–356.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika*, *53*(3), 349–359. https://link.springer.com/article/10.1007/BF02294217

Sinharay, S. (2003). *Assessing convergence of the Markov Chain Monte Carlo Algorithms: A review* (Research Report No. RR-03-07). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01899.x

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *64*, 583–616. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., & Best, N. (2003a). *WinBUGS* (Version 1.4) [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003b). *WinBUGS user manual*. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs

Sternberg, R. J. (1977). *Information processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*, 175–191. Retrieved from https://www.jstor.org/stable/1164643?seq=1#page_scan_tab_contents

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349–364. https://doi.org/10.1007/BF02295598

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, *51*, 589–601. https://doi.org/10.1007/BF02295598

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education*, *8*(2), 157–187.

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

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement*, *26*(1), 109–128. https://doi.org/10.1177/0146621602026001007

Wang, W., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318. https://doi.org/10.1177/0146621605276281

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

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Chicago: Scientific Software International.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Suggested citation:

Li, F. (2017). *An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory* (Research Report No. RR-17-27). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12151

Action Editor: Shelby Haberman

Reviewers: Tsung-Han Ho and Jianbin Fu