Journal of Educational Measurement
Summer 2006, Vol. 43, No. 2, pp. 145–168

Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests

Christine E. DeMars
James Madison University

Four item response theory (IRT) models were compared using data from tests where multiple items were grouped into testlets focused on a common stimulus. In the bi-factor model each item was treated as a function of a primary trait plus a nuisance trait due to the testlet; in the testlet-effects model the slopes in the direction of the testlet traits were constrained within each testlet to be proportional to the slope in the direction of the primary trait; in the polytomous model the item scores were summed into a single score for each testlet; and in the independent-items model the testlet structure was ignored. Using the simulated data, reliability was overestimated somewhat by the independent-items model when the items were not independent within testlets. Under these nonindependent conditions, the independent-items model also yielded greater root mean square error (RMSE) for item difficulty and underestimated the item slopes. When the items within testlets were instead generated to be independent, the bi-factor model yielded somewhat higher RMSE in difficulty and slope. Similar differences between the models were illustrated with real data.

Test items are often grouped into clusters, or *testlets*, centered around a common stimulus. For example, the items in a testlet may focus on a reading passage, a laboratory scenario, or a graphic or complex problem. Wainer, Bradlow, and Du (2000) describe several reasons testlets might be desirable. One of their reasons "is to reduce concerns about the atomistic nature of single independent small items" (p. 246). Testlets allow for more complicated, interrelated sets of items. Another reason they suggest is the efficient use of the examinee's time; if examinees must read a passage or study a stimulus, it is more time efficient to ask several related questions. These context-dependent items are often regarded as more realistic and perhaps even better for measuring higher-level skills. This may be particularly important in the current U.S. testing context; given the political prominence of and school time devoted to statewide testing, many educators desire tests that measure problem-solving in a context that is difficult to develop in a single item. However, items within a testlet often violate the item response theory (IRT) assumption of local independence (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987; Wang & Wilson, 2005; Yen, 1993); even after controlling for the primary trait measured by the test, the item response probabilities are not independent. Responses to items within a testlet, then, are related to a secondary trait. The secondary trait may be background knowledge or skills specific to the testlet, or it may be an interest level or other motivational factors specific to the testlet. Typically the testlet traits would be regarded as nuisance factors. While these testlet traits may be very important for a given problem, they do not generalize across contexts. The test user is not interested in estimating scores on these secondary traits, but ignoring them can lead to violations of the assumption of local independence.

Modeling Testlet Dependencies

One approach to this situation is to estimate a unidimensional model but treat items within a testlet as a single polytomous item (Cook, Dodd, & Fitzpatrick, 1999; Sireci et al., 1991; Wainer, 1995; Yen, 1993). The item scores are typically summed and the test is calibrated with an IRT model appropriate for polytomous items, such as the generalized partial credit model, the graded response model, or the nominal response model. Summing across items within the testlet, though, leads to some loss of information (Wainer et al., 2000). If there are a small number of items within the testlet the nominal model could be used with each response pattern coded as a different testlet score. This avoids the loss of information that would result from summing the item scores, but becomes impractical as the number of possible response patterns increases geometrically with the number of items in a testlet and thus is not frequently used (Thissen et al., 1989).

Responses to testlet items can also be modeled with a multidimensional model. The bi-factor model is appropriate for this context. In the bi-factor model, each item response is a function of the primary trait and one of the secondary traits. The secondary traits are orthogonal to the primary trait and to each other (Gibbons & Hedeker, 1992; McLeod, Swygert, & Thissen, 2001). When applied to testlets, the secondary traits would be testlet traits. The multidimensional extension of the three-parameter logistic (3PL) model is

$$P_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\frac{e^{1.7(\boldsymbol{a}_i'\boldsymbol{\theta}+d_i)}}{1 + e^{1.7(\boldsymbol{a}_i'\boldsymbol{\theta}+d_i)}}, \tag{1}$$

where $P_i(\boldsymbol{\theta})$ is the probability of correct response on item i given the $\boldsymbol{\theta}$ vector of traits and the item parameters, c_i is the lower asymptote, $\boldsymbol{a}_i$ is a vector of discrimination parameters, and d_i is the item difficulty. Notice that using this notation, d_i is added, not subtracted, so easier items have higher values for d in contrast to the usual unidimensional notation. For the bi-factor model, each item will have a nonzero value for a_{1i}, the discrimination in the direction of the primary trait, and for one other element of $\boldsymbol{a}_i$ corresponding to one of the testlet traits. The other discriminations in the vector, corresponding to the other testlets, are fixed to zero. The covariances among the traits are also fixed to zero. For identification purposes, the mean and variance of each θ would typically be set to zero and one, respectively. The TESTFACT (Bock et al., 2003) software has procedures to estimate the item discriminations and difficulties for the bi-factor model using marginal maximum likelihood (MML). TESTFACT uses the normal equivalent to the logistic model in Equation (1). TESTFACT will provide expected a posteriori (EAP) estimates of the primary trait, along with standard errors, which take the testlet structure into account.

Another model proposed for use with testlets, the testlet-effects model, involves a random testlet effect (Bradlow, Wainer, & Wang, 1999; Wainer et al., 2000; Wainer & Wang, 2000). The model is (Wainer & Wang, p. 205):

$$P = c_i + (1 - c_i)\frac{e^{1.7(a_i(\theta - b_i - \gamma_{g(i)}))}}{1 + e^{1.7(a_i(\theta - b_i - \gamma_{g(i)}))}}, \tag{2}$$

where P is the probability of correct response on item i by person j, c_i is the lower asymptote, a_i is the discrimination parameter, θ is the primary trait of person j, b_i is the item difficulty, and $\gamma_{g(i)}$ is the random testlet effect or testlet trait for person j of testlet $g(i)$, the testlet to which item i belongs. The variance of $\gamma_{g(i)}$ is a parameter to be estimated; larger variance indicates a greater effect for testlet g. When multiplied by a_i, the difficulty b_i in Equation (2) should be the negative of the difficulty d_i in Equation (1). This model was originally specified by Bradlow et al. without the lower asymptote and without the 1.7 constant.

Li, Bolt, and Fu (2004) observed that because the testlet-effects model applies the same discrimination parameter to both the primary trait θ and the testlet trait γ, items that discriminate well on the primary trait are also modeled as discriminating well on the testlet trait, when the opposite might seem more reasonable. They modified the model to include separate discrimination parameters for the primary and testlet traits, equivalent to the bi-factor model, and found that this model fit their data better. The relationship with the bi-factor model can be illustrated by re-writing Equation (2) as follows:

$$P = c_i + (1 - c_i)\frac{e^{1.7(a_i\theta - a_i b_i - a_i\gamma_{g(i)})}}{1 + e^{1.7(a_i\theta - a_i b_i - a_i\gamma_{g(i)})}}, \tag{3}$$

where all terms are as defined for Equation (2). If $d_i = -a_i b_i$, and $\gamma_{g(i)}$ is scaled to have a mean of 0 and *SD* of 1 and symbolized as $\theta_{g(i)+1}$, the equation becomes

$$P = c_i + (1 - c_i)\frac{e^{1.7(a_i\theta_1 - a_i\alpha_{g(i)}\theta_{g(i)+1} + d_i)}}{1 + e^{1.7(a_i\theta_1 - a_i\alpha_{g(i)}\theta_{g(i)+1} + d_i)}}, \tag{4}$$

where $\alpha_{g(i)}$ is equal to the *SD* of $\gamma_{g(i)}$ in Equation (2). Comparing Equation (1) to Equation (4), the difference is that the testlet slope in Equation (4) is a product of the item slope, a_i, (a_{1i} in Equation (1)) and a testlet constant, $\alpha_{g(i)}$, that is equal for all items within the same testlet. If the product were written as a single slope as in Equation (1), the testlet slopes within the same testlet would be proportional to the primary slope. In the bi-factor model, the testlet slopes are independent of the primary slope. Thus, the testlet model is a constrained version of the bi-factor model. Li et al. (2004) showed this relationship, except with no lower asymptote in the model.

The standard 3PL model for independent items is nested within the testlet-effects (as well as the bi-factor) model; γ is constrained to zero and thus it and the corresponding slope drop out of the model. Also, the model is typically transformed such that the item difficulty, *b*, is equal to $-d/a$ and *a* is brought outside the parentheses.

Consequences of Ignoring Local Dependencies

Ignoring violations of local independence can lead to overestimates of reliability or information and underestimates of the standard error of the ability estimates

(Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). It can also lead to misestimation of item parameters. Wainer and Wang (2000) found that when testlet dependencies were ignored item difficulties were still well estimated but lower asymptotes were overestimated. Discriminations were underestimated on one test and overestimated on another. Bradlow, Wainer, and Wang (1999) showed that when the testlet dependencies were not modeled the item discriminations for testlet items were underestimated and the item discriminations for independent items were overestimated. Ackerman (1987) found that item discriminations were overestimated when a subset of the items were locally dependent. Wainer et al. (2000) also showed that traits and difficulties were recovered better than discriminations and lower asymptotes when the testlet dependencies were omitted from the model. Further, the item discriminations from the testlet model replicated in another sample much better than the item discriminations from the usual 3PL model. In a study by Glas, Wainer, and Bradlow (2000), the mean absolute errors of both discrimination and difficulty were larger for the 3PL model than for the testlet model when the testlet factor was large. Lee, Kolen, Frisbie, and Ankenmann (2001) found that treating each testlet as a polytomous item was more effective in equating than ignoring the local dependencies and using the usual 3PL model, which suggests that ignoring the dependency led to less accurate item parameter estimates. Dresher (2004) found that, when all items on a test were in testlets, the root mean square error (RMSEs) for ability were higher when the dependencies were ignored than they were when the testlets were each treated as a polytomous-cluster or when the testlet-effects model was used. For the item difficulties and discriminations, RMSEs were lowest for the testlet-effects model.

Purpose

Most of the studies summarized above have compared either the independent-items model with the testlets-as-polytomous-item model (Lee et al., 2001; Sireci et al., 1991; Wainer, 1995; Yen, 1993) or the independent-items model with the testlet-effects model (Bradlow et al., 1999; Glas et al., 2000; Wainer et al., 2000; Wainer & Wang, 2000). Li et al. (2004) compared the equivalent of the bi-factor model to the testlet-effects model, but not to the polytomous model or the independent-items model. In the current study, item and trait parameters for simulated data and for two real testlet-based tests were estimated using the bi-factor model, the testlet-effects model, the testlets-as-polytomous-items model, and the independent-items model. The purpose was to compare the ability, reliability, item difficulty, and item discrimination estimates from the different models. These comparisons were made with data generated to fit the bi-factor model, with data generated to fit the testlet-effects model, and again with data generated to fit the independent-items model. Because the independent-items model is nested within the testlet-effects and bi-factor model, the more complex models should lead to the same average item and trait estimates with data generated from the independent-items model but they may introduce more error due to chance overfitting. Similarly, because the testlet-effects model is nested within the bi-factor, the bi-factor model may lead to less stable parameter estimates when the data follow the testlet-effects model. When the model used for parameter estimation is more constrained than the model used to generate the data (independent-items model used with testlet-effects or bi-factor data, or testlet-effects used with bi-factor data), item or trait estimates may be inaccurate, as several researchers have found when using the independent-items model when items were not locally independent (Ackerman, 1987; Bradlow et al., 1999; Glas et al., 2000; Wainer et al., 2000; Wainer & Wang, 2000). Also in the latter conditions, estimated standard errors would generally be too low, and reliability estimates would be too high (Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). Finally, using the testlets-as-polytomous items model the local dependencies from the bi-factor and testlet-effects should not systematically affect the trait estimates or estimated reliability (item parameter estimates will not be comparable). However, the loss of information from summing items would be expected to lead to true increases in RMSE and decreases in reliability.

Study 1: Simulation Study

Method

Data. For consistency, the parameters (slopes and item difficulties) used to generate the data are described in terms of the parameters of Equation (1). Six tests were created by crossing two test lengths (25 or 50 items) with three models (bi-factor model, testlet-effects model, and unidimensional model). Each set of five items formed a testlet, with one of five magnitudes of testlet effects. Primary item slopes ranged from .6 to 1.4 and item difficulties ranged from -1.5 to 1.5. Lower asymptotes were set to .2. The same primary slopes and difficulties were used with each of the three simulation models. When the testlet slopes were independent of the primary slope, as in the bi-factor model, the testlet slopes were set to 0 in one testlet, and .3, .6, .9, and 1.2 in each of the others. The testlet slope was set the same within each testlet only for convenience in organizing results and to keep the testlet and primary slopes independent; these slopes were not constrained to be equal when estimating the bi-factor model. When the testlet slopes were proportional to the primary slope, as in the testlet-effects model, the testlet slopes were set to 0 in one testlet, and .3 times the primary slope in another testlet, and .6, .9, or 1.2 times the primary slope in each of the others. For the unidimensional model, all the testlet slopes were 0. The item parameters used to generate the data are shown in Table 1; for the tests with 25 items the first 25 items in the table were used.

For each of the conditions, 2,000 simulated response patterns were generated. For each simulee, a primary trait and 10 testlet traits were independently drawn from standard normal distributions. Based on the item parameters, the primary trait, and the appropriate testlet trait, probability of correct response was calculated for each simulee on each item and if a draw from a uniform distribution between 0 and 1 was less than this probability the response was coded correct. This process was repeated 100 times. The same data sets were used for the 25-item and 50-item conditions, except that only the first 25 items were used in the 25-item condition.

Estimation. The item parameter and primary traits for all conditions, regardless of the model used to generate the item parameters, were estimated using the bi-factor model, the unidimensional 3PL model, the testlet-effects model, and the testlets-as-polytomous-item model.

TABLE 1
Item Parameters Used to Simulate the Data

			Testlet Slope	
Item	Primary Slope	Difficulty	Bi-Factor	Testlet-Effects
1	0.6	−1.5	0.0	0.00
2	0.8	−0.5	0.0	0.00
3	1.0	0.0	0.0	0.00
4	1.2	0.5	0.0	0.00
5	1.4	1.5	0.0	0.00
6	0.6	−0.5	0.3	0.18
7	0.8	0.0	0.3	0.24
8	1.0	0.5	0.3	0.30
9	1.2	1.5	0.3	0.36
10	1.4	−1.5	0.3	0.42
11	0.6	0.0	0.6	0.36
12	0.8	0.5	0.6	0.48
13	1.0	1.5	0.6	0.60
14	1.2	−1.5	0.6	0.72
15	1.4	−0.5	0.6	0.84
16	0.6	0.5	0.9	0.54
17	0.8	1.5	0.9	0.72
18	1.0	−1.5	0.9	0.90
19	1.2	−0.5	0.9	1.08
20	1.4	0.0	0.9	1.26
21	0.6	1.5	1.2	0.72
22	0.8	−1.5	1.2	0.96
23	1.0	−0.5	1.2	1.20
24	1.2	0.0	1.2	1.44
25	1.4	0.5	1.2	1.68
26	0.6	1.5	0.0	0.00
27	0.8	0.5	0.0	0.00
28	1.0	0.0	0.0	0.00
29	1.2	−0.5	0.0	0.00
30	1.4	−1.5	0.0	0.00
31	0.6	0.5	0.3	0.18
32	0.8	0.0	0.3	0.24
33	1.0	−0.5	0.3	0.30
34	1.2	−1.5	0.3	0.36
35	1.4	1.5	0.3	0.42
36	0.6	0.0	0.6	0.36
37	0.8	−0.5	0.6	0.48
38	1.0	−1.5	0.6	0.60
39	1.2	1.5	0.6	0.72
40	1.4	0.5	0.6	0.84
41	0.6	−0.5	0.9	0.54
42	0.8	−1.5	0.9	0.72
43	1.0	1.5	0.9	0.90
44	1.2	0.5	0.9	1.08
45	1.4	0.0	0.9	1.26
46	0.6	−1.5	1.2	0.72
47	0.8	1.5	1.2	0.96
48	1.0	0.5	1.2	1.20
49	1.2	0.0	1.2	1.44
50	1.4	−0.5	1.2	1.68

The bi-factor model was estimated using TESTFACT. All items were free to load on the primary factor, and items from the same testlet were specified as loading on an additional secondary factor. The lower asymptote was fixed to .2. The maximum number of cycles was increased to 50, which generally led to a maximum change of .01 between the last two iterations. For item estimation, 19 quadrature points were used for each trait in the 25-item tests and 9 quadrature points were used for each trait in the 50-item tests, with default priors applied to the slopes. The score of interest in this context was the primary trait, estimated in TESTFACT using EAP scoring with a normal prior and nine quadrature points.

The independent-items model was also estimated using TESTFACT, using the same options, but with all items loading only on the primary factor.

The testlet-effects model was estimated using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). The model was specified in terms of Equation (4) because it could be easily transformed to the parameters in Equation (1) for comparisons. The prior distributions for the primary and testlet traits were each standard normal and limited to the interval (−5, 5). The primary slope and testlet coefficient each had a log-normal prior with a mean of zero and a precision of .25 (*SD* of 2). The prior for the item difficulties was $N(0, 2)$, limited to the interval (−4, 4). Initial trait values and item difficulties were all set to zero; initial slope values were set to one. For consistency with the bi-factor and unidimensional models, the lower asymptotes were fixed to .2. The first 1,000 iterations were treated as the burn-in and were discarded; this was conservative and convergence generally appeared to be reached in 200–300 iterations. Estimates were obtained from the mean of the next 2,000 iterations. Because of the time needed to run this procedure, only the first 50 replications were run for each condition.

PARSCALE (Muraki & Bock, 2003) was used for the testlet-as-polytomous-item model. The item scores in each testlet were summed and the sum was treated as a single item using the generalized partial credit model (Muraki, 1992). Default priors were applied to the slopes and step parameters, and the POSTERIOR option was used to adjust the scaling such that the estimated posterior distribution had a mean of zero and *SD* of one. The default 30 quadrature points were used. As with the other three models, traits were estimated using EAP scoring and a standard normal prior.

TESTFACT and PARSCALE use MML procedures; to avoid possible confounding of model with estimation method, the bi-factor, unidimensional, and testlet-as-polytomous-item models could have been specified in WinBUGS. However, one of the purposes of this paper was to compare these models as they are commonly used in practice. The MML procedures in TESTFACT and PARSCALE require a small fraction of the time that MCMC procedures require to run, and they require a far lower level of user knowledge, so these software packages are in widespread use. Similarly, one could write a program to estimate the testlet-effects model using MML (see Glas, Wainer, & Bradlow, 2000 for an example), but WinBUGS or self-programmed MCMC procedures are more commonly used.

Results

Model fit. Because the testlet-effects model is nested within the bi-factor model (see Equations (1)–(4)) and the independent-items model is nested within the testlet-effects model, the significance of the difference in -2 log-likelihood can be tested with a χ^2-difference test[1] (du Toit, 2003, pp. 587–588). TESTFACT provides the marginal -2 log-likelihood, the probability of the data averaged over the trait distribution. One advantage of using the marginal likelihood is that it does not depend on the accuracy of the trait estimates; studies with item-level fit measures for unidimensional items have shown that using trait estimates changes the distribution of the fit index (Orlando & Thissen, 2000; Orlando & Thissen, 2003; Stone, 2000; Stone, 2003; Stone & Hansen, 2000). Given that the testlet traits are each based on only five items and thus are likely to be unstable, their use might be particularly problematic. Marginalizing over the trait distribution avoids using these trait estimates. The deviance statistic reported in WinBUGS is *not* marginalized over the trait distribution, so the marginal -2 log-likelihood for the testlet-effects model was calculated separately, using a routine written in SAS. To avoid differences due to the number of quadrature points or to rounding differences in the algorithm, this procedure was also used for the bi-factor and independent-items model; in trial runs using the quadratures used by TESTFACT, the results were nearly identical to those reported in the software output. Quadrature points from -3 to 3 at intervals of .5 were used for each trait. For each examinee, the probability of the testlet response pattern was found for each combination of the primary trait and the testlet trait, then marginalized across the testlet trait distribution. The probability of the complete response pattern was then found at each quadrature of the primary trait, and marginalized across the primary trait distribution. The model marginal -2 log-likelihood was then calculated as the negative of twice the sum, over examinees, of the natural logs of these likelihoods. Fit was calculated only for the first 50 replications within each condition because of the extensive time involved; this was adequate to give a picture of whether the models fit as expected, and the purpose was not to estimate exact Type I error rates or power, which would have required many more replications. The fit of the testlets-as-polytomous-items model was not compared to the fit of the other models. Even if an index appropriate for comparing nonnested models such as the Akaike Information Criterion (AIC) were used, the comparison would not be meaningful because, for the testlets-as-polytomous-items model, the likelihood would have to be computed for the summed testlet score, not the individual items.

As expected, the more complex model fit better than the constrained models, but this difference was most often not significant when the data were generated using the constrained model. The only unexpected result was when the data were generated by the independent-items model and there were 50 items; the testlet-effects model always fit the data significantly better in this condition. The ratio of the likelihoods was quite close, .9996 on average, yet the difference was statistically significant. This suggests there may be some problems in using this index with large numbers of items. With this exception, the more complex model fit better than the constrained model when the data were generated with a more complex model but not when the data were generated by the more constrained model.

Accuracy of trait estimates. Before comparing the trait estimates, if needed the item and trait parameters for each replication were scaled such that the estimated population posterior distribution for the primary trait had a mean of zero and *SD* of one.

This rescaling was based on the estimated posterior distribution, *not* on the distribution of the estimated scores. The mean of the estimated traits is approximately equal to the mean of the posterior trait distribution, but the *SD* of the estimated scores underestimates the *SD* of the trait distribution, because the EAP scores are biased toward the mean (du Toit, 2003, p. 607). Rescaling such that the posterior distribution had a mean of zero and *SD* of one was done within the PARSCALE routine for the testlets-as-polytomous-items model by choosing the POSTERIOR option. In the MCMC routine used for the testlet-effects model, setting the prior distribution of the primary trait to a *SD* of one also resulted in an estimated posterior distribution with a *SD* of one. For the 25-item tests the posterior distribution also seemed to be scaled to these constants in TESTFACT as well; the same was not true for the 50-item tests, where the posterior *SD* was generally greater than one. The sum of the variance of the estimated traits and the average error variance equals the variance of the posterior distribution printed after the last iteration in the TESTFACT output, so this sum was used in rescaling the items and primary traits. For consistency, rescaling was applied to all models, though the constants were virtually zero and one in all conditions except for the 50-item tests estimated with the bi-factor and independent-items models.

For each simulee, within each condition, bias was calculated using the typical definition of the mean difference across the 100 replications (50 replications for the testlet-effects model) between the estimated primary trait and the true primary θ used to generate the data. RMSE for each simulee was calculated as the square root of the average squared difference between the estimated and true primary θ. These values would be expected to vary as a function of θ, so they are plotted across the θ distribution in Figures 1 and 2 for the conditions in which the data were simulated using the bi-factor model and 25 items. The same patterns held in the other conditions, except where noted. In the middle ranges of θ, bias and RMSE were nearly equal regardless of the estimation model. At the extreme ends, where bias and RMSE were greater in general, they were greater for the testlet-as-polytomous-item model than for the other models, likely because the slight loss of information from using this model gave the prior distribution greater weight. These patterns were similar for the conditions where the data were generated with the other models (not shown to save space, available on request) and for the 50-item tests, except that for the 50-item tests, at high levels of θ the testlet-as-polytomous-item model did not yield greater bias and RMSE than the other models. As would be expected, the RMSE and the absolute value of bias were lower when there were 50 items instead of 25. RMSE was slightly but consistently lower when the data were generated with the independent-items model. This same information was averaged across the primary θ and displayed in Table 2. Only the mean RMSE and not the mean bias is not shown because in all conditions bias at one end balanced bias at the other and the mean bias was virtually zero.

Because RMSE is the deviation around the true value of the parameter, it includes the effects of bias. The deviation around the expected value does not include the bias. The *SD* around the average estimated θ is shown in Figure 3, again for the conditions in which the data were simulated using the bi-factor model and 25 items. These average estimates spanned a narrower range than the true θs because the extreme values were biased inward. The *SD*s were fairly stable across the range of

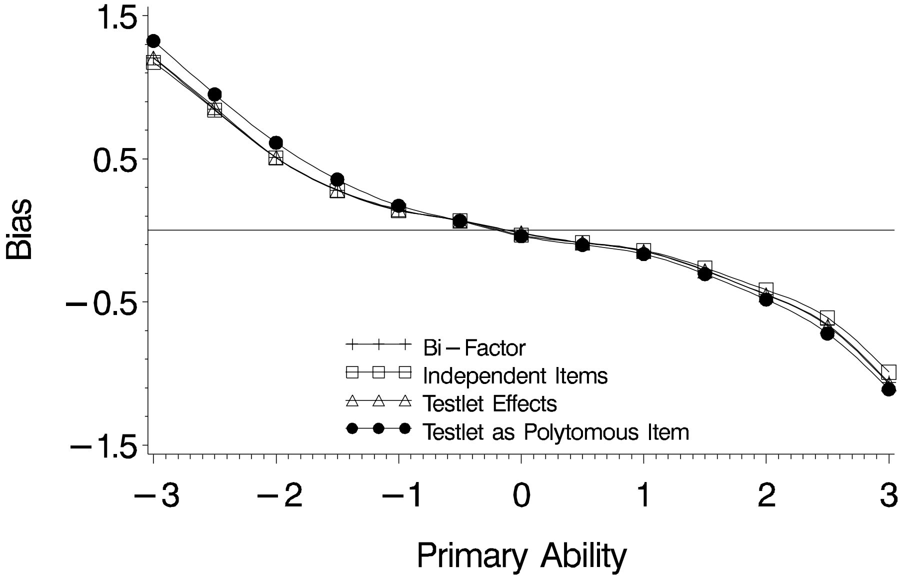

FIGURE 1. *Bias for the primary θ, for the 25-item test with the data generated using the bi-factor model.*

estimated θ. Without the prior distribution, the *SD*s would have been higher at the extremes, but the prior distribution had a larger effect at the extremes, and as there was no variance in the prior this decreased the variance in the estimated abilities.

Reliability and estimated reliability. One definition of reliability in the classical test theory is the squared correlation between the estimated and true scores. The true θs were known because the data were simulated. The correlation was calculated within each replication and averaged over replications, with the results reported in Table 3. Within the same simulation model and test length, these reliabilities were very similar regardless of the model used to estimate the primary θ; reliability was very slightly lower for the polytomous model due to a small loss of information from ignoring the exact pattern of item responses within the testlet (as suggested by Wainer & Wang, 2000; Zenisky, Hambleton, & Sireci, 2002). Comparing the different data-generating models, reliability was higher for the independent-items model because the testlet θ added additional error from the perspective of estimating the primary θ. This is consistent with the finding of smaller RMSE for this model in Table 2.

Usually the true θs are not known and reliability must be estimated. In IRT, where the standard error is not constant across examinees, marginal reliability can be estimated as $1 - (s_\mathrm{e}^2/s_\mathrm{T}^2)$, where s_e is the average standard error and s_T is the estimated *SD* of the population distribution (the sum of the variance of the EAP scores and the error variance). This reliability estimate is labeled the empirical reliability in TESTFACT (du Toit, 2003, p. 34). Again, this parallels classical test theory, except that s_e is not a constant across examinees. These reliability estimates are also reported in Table 3. When the data were generated using the bi-factor or testlet-effects

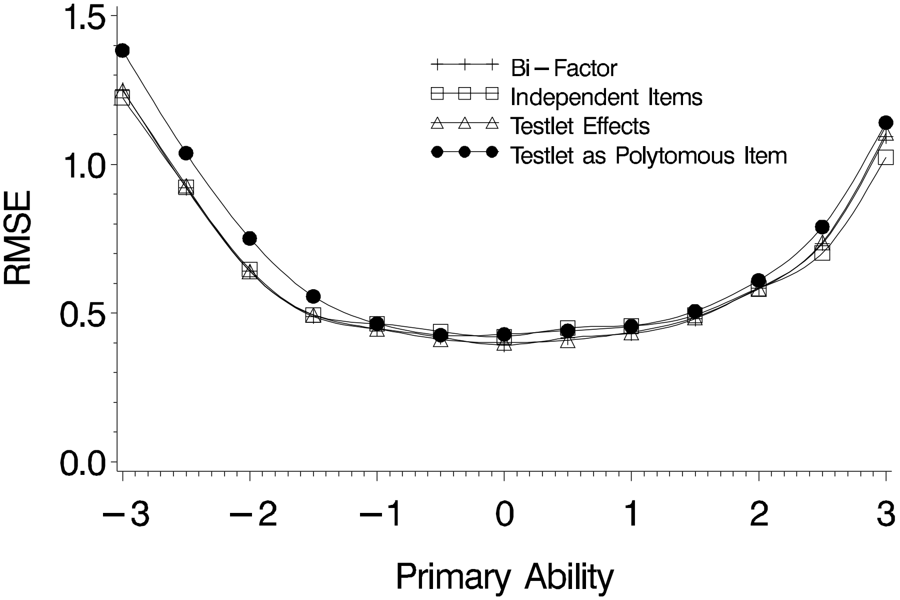

FIGURE 2. *RMSE for the primary θ, for the 25-item test with the data generated using the bi-factor model.*

model, the reliability estimates were higher when the scores were estimated using the independent-items model. These scores were not truly more reliable. Rather, the independent-items model overestimated the reliability of the scores when the data were generated by the bi-factor or testlet-effects model. These scores appeared to be more precise than they really were.

Accuracy of estimated item parameters. Differences in the item parameters matter not only because they can impact the trait scores but also because they may be used

TABLE 2
Mean RMSE of the Primary θ Scores

		Data Simulation Model		
Test Length	Estimation Model	Bi-Factor	Testlet-Effects	Independent Items
25				
	Bi-Factor	.21	.22	.17
	Testlet-Effects	.21	.22	.16
	Independent Items	.23	.23	.17
	Polytomous Items	.25	.24	.19
50				
	Bi-Factor	.13	.13	.10
	Testlet-Effects	.12	.12	.09
	Independent Items	.14	.14	.10
	Polytomous Items	.14	.14	.11

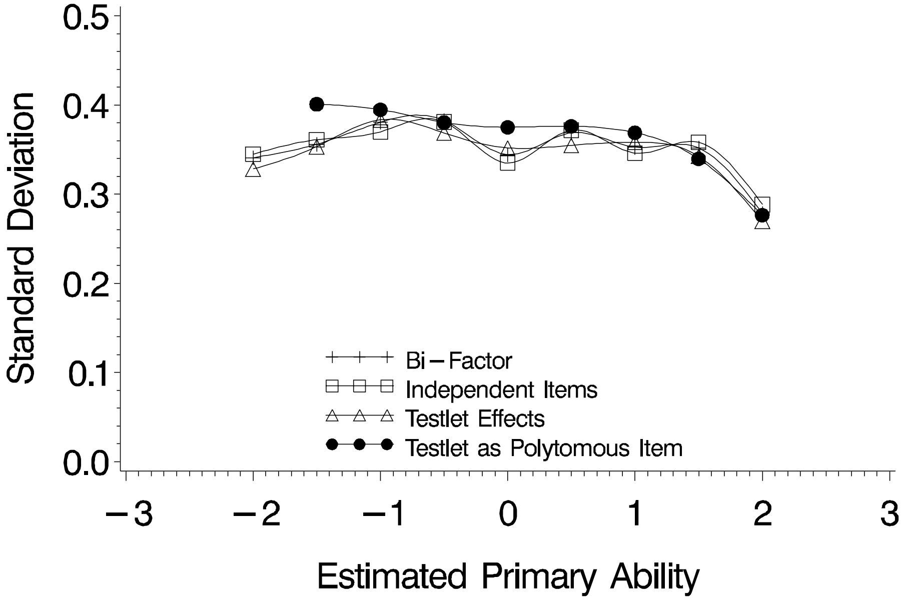

FIGURE 3. *SD of the estimated primary θ, for the 25-item test with the data generated using the bi-factor model.*

to equate test forms or to select items for a test form. Because the testlet-effects and independent-items models are of the same form as the bi-factor model, with additional constraints, the item parameters from these models can be reasonably compared. The testlets-as-polytomous-items model is of a different form and so it will not be included in these comparisons.

TABLE 3
Mean Squared Correlation Between True Primary θ and Estimated Primary θ, and Model-Based Estimated Reliability of the Primary θ Scores

		Data Simulation Model		
Test Length	Estimation Model	Bi-Factor	Testlet-Effects	Independent Items
25				
	Bi-Factor	.792, .795	.787, .790	.837, .845
	Testlet-Effects	.792, .793	.791, .788	.846, .843
	Independent Items	.775, .827	.773, .823	.837, .846
	Polytomous Items	.761, .765	.762, .765	.817, .824
50				
	Bi-Factor	.873, .883	.871, .880	.898, .913
	Testlet-Effects	.882, .882	.881, .879	.915, .912
	Independent Items	.864, .901	.862, .899	.898, .915
	Polytomous Items	.860, .866	.861, .867	.892, .903

Note. The first value in each cell is the mean correlation between true and estimated scores, and the second value is the model-based reliability estimate.

Table 4 shows the mean difference between the estimated and true parameters and RMSE for item difficulties, in the conditions where the data followed the bi-factor model. Because the bias was averaged over items within each testlet, bias toward the mean would be compensated for; only bias consistently in the same direction would appear in the mean difference. Within each condition, results are broken down by the size of the testlet slope for the bi-factor model or the testlet coefficient for the testlet-effects model (the coefficient $\alpha_{g(i)}$ in Equation (4), which is multiplied by the primary slope to obtain the testlet slope in Equation (1)). None of the estimation models led to consistent under- or over-estimation of the item difficulties, though the RMSE was larger for the independent items when the testlet slope was large, and larger for the bi-factor model when the testlet slope was zero. The same pattern was seen in Table 5 where the data were generated using the testlet-effects model. Finally, in Table 6 where the data were generated using the independent-items model, the RMSE was greatest for the bi-factor model.

Mean differences and RMSE for the slopes in the direction of the primary trait are shown in Tables 7–9. Unlike the item difficulties, the slopes were consistently underestimated by the independent-items model when the data were generated with the bi-factor or testlet-effects models. The extent of the underestimation increased as the testlet slope increased. The testlet-effects model, when used with data that followed the bi-factor model, overestimated the slopes slightly and had increased RMSE when the testlet slope was large. In contrast, the bias and RMSE for the bi-factor model, when used with data that followed the testlet-effects model, was no larger and was sometimes smaller than that of the testlet-effects model. When the data were generated by the independent-items model, using the bi-factor or testlet-effects models for estimation led to a small positive bias, and the bi-factor model increased the RMSE of the slopes.

Real Data Example

In the simulation study, data were generated to fit at least one of the models. Real data generally will not fit any of the models as well. This second study is an example showing how the model fit, primary trait estimates, standard errors/reliability, and item parameters compare when the four models are used with two real data sets.

Method

Instrument. The data sets used in this study were from the Programme for International Student Assessment 2000 (PISA 2000) public release data set (Organization for Economic Cooperation and Development, 2002). Selected math and reading tests were utilized for this study.

The items on the math test came from Booklet 1. There were 15 items: two 2-item testlets, two 3-item testlets, one 4-item testlet, and one independent item. Each item was scored right or wrong.

The items on the reading test came from Booklet 7 of the PISA 2000 data set. In the PISA study, three types of reading items were used: retrieving information, interpreting texts, and reflection and evaluation. The PISA technical report (Adams & Wu,

TABLE 4
Mean Difference Between True and Estimated Item Difficulties and RMSE, Data Generated with the Bi-Factor Model

		Testlet Slope									
		0		.3		.6		.9		1.2	
Test Length	Estimation Model	MD[a]	RMSE	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE
25											
	Bi-Factor	−.03	.15	.00	.11	−.02	.12	.06	.13	.03	.14
	Testlet-Effects	.01	.09	.00	.12	−.04	.12	.04	.12	.03	.17
	Independent Items	−.02	.10	.01	.13	.00	.17	.04	.25	.03	.34
50											
	Bi-Factor	−.03	.15	−.01	.11	−.01	.13	.04	.14	.01	.14
	Testlet-Effects	.01	.08	.00	.10	.00	.13	.04	.12	.00	.19
	Independent Items	.00	.10	.00	.12	.00	.18	.03	.27	.00	.36

[a] Mean difference.

TABLE 5

Mean Difference Between True and Estimated Item Difficulties and RMSE, Data Generated with the Testlet-Effects Model

		Testlet Coefficient									
		0		.3		.6		.9		1.2	
Test Length	Estimation Model	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE
25											
	Bi-Factor	−.04	.14	−.01	.12	−.02	.13	.05	.12	.03	.13
	Testlet-Effects	.00	.08	.01	.08	.00	.11	.05	.11	.03	.11
	Independent Items	−.02	.10	.02	.14	.02	.18	.08	.23	.04	.25
50											
	Bi-Factor	−.03	.14	−.01	.13	−.01	.13	.04	.13	.00	.13
	Testlet-Effects	−.01	.09	.01	.08	.01	.10	.04	.11	.00	.11
	Independent Items	.00	.10	.00	.13	.00	.19	.03	.24	.01	.27

TABLE 6
Mean Difference Between True and Estimated Item Difficulties and RMSE, Data Generated with the Independent-Items Model

Test Length	Estimation Model	MD	RMSE
25			
	Bi-Factor	−.03	.14
	Testlet-Effects	.00	.08
	Independent Items	−.01	.09
50			
	Bi-Factor	−.01	.15
	Testlet-Effects	.00	.09
	Independent Items	.01	.10

2002, p. 153) reported a correlation of .97 between the retrieving and interpreting factors, so for the current study both types of items were treated as a unidimensional construct; the reflection and evaluation items were not used because the model would have been much more complex if they were included: two correlated primary traits, each with multiple testlet traits. This left 40 items in 14 testlets: 8 two-item testlets, 3 three-item testlets, and 3 five-item testlets. In the PISA database manual, 38 of these items were scored right or wrong while the scoring for two items allowed for partial credit. For the current study, the partial credit scores were coded as incorrect so that dichotomous IRT models could be used.

Participants. For each test, 5,000 examinees were randomly selected from among those who *completed* the selected test booklet. Students who left items blank at the end of the test were considered noncompleters and were not eligible for selection for the current study, but students who omitted items in the middle of the test were eligible for selection. Items omitted in the middle of the test were scored as incorrect.

Estimation. The models were estimated using the same software programs and options as were used for the simulation study.

Local-dependence due to testlets. DIMTEST (Stout, Douglas, Junker, & Roussos, 1999) was used to test for essential unidimensionality; if the tests were unidimensional then there would be no significant testlet factors. In DIMTEST, a group of items possibly measuring a secondary dimension is designated the assessment test. Another group of items of similar difficulty is reserved to correct for bias, and examinees are grouped by their scores on the remaining items, called the partitioning test. Information on calculating the test statistic can be found in the DIMTEST manual. For the present study, the items from one testlet that accounted for the greatest proportion of variance in the bi-factor analysis were used as the assessment test. For both the reading and math tests, Stout's T was statistically significant ($p < .0001$), indicating the tests were not essentially unidimensional.

TABLE 7

Mean Difference Between True and Estimated Slopes and RMSE, Data Generated with the Bi-Factor Model

		Testlet Slope									
		0		.3		.6		.9		1.2	
Test Length	Estimation Model	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE
25											
	Bi-Factor	.03	.13	.01	.13	.00	.13	−.02	.12	.00	.12
	Testlet-Effects	.02	.11	.03	.14	.06	.20	.05	.18	.06	.15
	Independent Items	−.08	.13	−.09	.14	−.12	.16	−.17	.20	−.19	.21
50											
	Bi-Factor	.02	.13	−.01	.11	−.01	.12	−.02	.12	−.02	.12
	Testlet-Effects	.02	.10	.02	.11	.07	.20	.07	.21	.06	.19
	Independent Items	−.06	.11	−.09	.13	−.15	.17	−.23	.26	−.31	.33

TABLE 8
Mean Difference Between True and Estimated Slopes and RMSE, Data Generated with the Testlet-Effects Model

		Testlet Coefficient									
		0		.3		.6		.9		1.2	
Test Length	Estimation Model	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE	MD	RMSE
25											
	Bi-Factor	.03	.12	.02	.13	.01	.15	−.03	.14	−.01	.15
	Testlet-Effects	.02	.10	.02	.10	.04	.16	.02	.15	.05	.18
	Independent Items	−.08	.13	−.09	.14	−.13	.18	−.20	.26	−.23	.29
50											
	Bi-Factor	.02	.13	.00	.12	.00	.14	−.02	.15	−.02	.14
	Testlet-Effects	.03	.11	.00	.10	.03	.13	.03	.15	.04	.17
	Independent Items	−.06	.11	−.09	.14	−.16	.20	−.25	.31	−.33	.39

TABLE 9

Mean Difference Between True and Estimated Slopes and RMSE, Data Generated with the Independent-Items Model

Test Length	Estimation Model	MD	RMSE
25			
	Bi-Factor	.04	.14
	Testlet-Effects	.03	.11
	Independent Items	.00	.10
50			
	Bi-Factor	.01	.12
	Testlet-Effects	.03	.10
	Independent Items	−.04	.10

Results

Using the difference in −2 log-likelihood test described in the simulation study, for the math test the bi-factor model fit better than the testlet-effects model ($\chi^2(9) = 670, p < .0001$), which in turn fit better than the independent-items model ($\chi^2(5) = 103, p < .0001$). Similarly, for the reading test the bi-factor model fit better than the testlet-effects model ($\chi^2(26) = 81, p < .0001$), which in turn fit better than the independent-items model ($\chi^2(14) = 1182, p < .0001$).

Because real data were used and the true parameters were unknown, estimates from the different methods were compared to each other rather than to the true parameters. For the reading test, as for the 50-item simulated tests, the estimated posterior distribution of the primary trait in TESTFACT had a *SD* greater than one so the trait and item estimates were rescaled using the procedures described for the simulated data. The average correlations among traits and the average root mean square difference (RMSD) between the traits are shown in Table 10. The RMSD was calculated in the same way as the RMSE would be, except that it was based on the difference between two estimates instead of the difference between an estimate and a true value. The RMSD were averaged across traits, rather than graphed by ability as was done in Figures 1 and 2, because the number of pairs to be compared would have yielded multiple figures (or many lines within each figure). Differences were of course greater further away from the mean.

Correlations among the trait estimates from the different models were very high (at least .99). The average difference between trait estimates was nearly zero in all cases, so no model consistently overestimated or underestimated ability relative to the other models; this would be expected because each method scales the estimates such that the estimated posterior distribution has a mean of zero and *SD* of one. Therefore, the average differences were not displayed in the table. In math, the RMSD was somewhat smaller between the bi-factor model and the testlet-effects model, and it was largest between the independent-items and each of the other three models. In reading the bi-factor and independent-items model estimates had the smallest RMSD.

Reliability was estimated as described for the simulation study, as the ratio of the variance of the trait estimates (EAP scores) to the sum of this variance plus the

TABLE 10

Correlations Among θ Estimates and Root Mean Square Difference in θ Estimates

	Correlation	RMSD
Math		
Bi-Factor vs. Testlet Effects	.999	.04
Bi-Factor vs. Independent Items	.992	.12
Bi-Factor vs. Polytomous Items	.995	.09
Testlet Effects vs. Independent Items	.990	.13
Testlet Effects vs. Polytomous Items	.995	.09
Independent Items vs. Polytomous Items	.990	.13
Reading		
Bi-Factor vs. Testlet Effects	.995	.09
Bi-Factor vs. Independent Items	.998	.05
Bi-Factor vs. Polytomous Items	.991	.12
Testlet Effects vs. Independent Items	.993	.11
Testlet Effects vs. Polytomous Items	.996	.08
Independent Items vs. Polytomous Items	.990	.14

average squared standard error, which is the *empirical reliability* in the TESTFACT output. The reliability estimates for math were .80 for the independent-items model and .75 for the other models. For reading, the estimates were .89 for the independent-items model and .87 for the other models. If the testlet effects are small, there should be little or no difference between the estimated reliability using the independent-items model and the estimated reliability using the other models. This seems to be the case for the reading test. If the testlet effects are large enough to make a difference in the reliability estimates, then the independent-items reliability is likely an overestimate, which appears to be the case for the math test.

Next, the item parameter estimates were compared. Table 11 shows the average difference and RMSD in item difficulty and for the item discriminations in the direction of the primary trait. The testlets-as-polytomous-items method was not included in these tables because the item parameters were not in a comparable form and cannot be transformed to a comparable form. The items had slightly higher difficulty estimates in the testlet-effects model, which means the items appeared easier (recall that the difficulty in Equation (1) is added, not subtracted, so higher difficulties indicate easier items). Item discriminations were lower for the independent-items model, as they were in the simulation study whenever there was nonindependence within the testlets.

Overall, the math test showed greater differences than the reading test between the independent-items model and the other models that explicitly took the testlet structure into account. These differences suggest that the nonindependence due to testlets was a larger problem on the math test. For the reading test, it made almost no difference which model was used, though there was a slight inflation in reliability and underestimation of item discrimination when the independent-items model was used.

TABLE 11
Mean Difference and Root Mean Square Difference in Item Parameter Estimates

	Difficulty		Slope	
	MD	RMSD	MD	RMSD
Math				
Bi-Factor vs. Testlet Effects	−.07	.28	.06	.18
Bi-Factor vs. Independent Items	.01	.46	.19	.39
Testlet Effects vs. Independent Items	.08	.33	.13	.27
Reading				
Bi-Factor vs. Testlet Effects	−.01	.16	.00	.10
Bi-Factor vs. Independent Items	.09	.27	.10	.18
Testlet Effects vs. Independent Items	.11	.20	.10	.15

Discussion and Conclusions

Trait Estimates

Reliability is defined in classical test theory as the squared correlation between the true and observed scores. Using this definition for the simulated data where the true scores were known, the reliability of the four sets of estimated scores was similar (slightly lower for the testlet-as-polytomous-items estimation model). These reliabilities were lower than those when there were no testlet effects. Nonindependence of items within testlets decreased the reliability of trait estimates, regardless of which model was used to estimate the traits, because the testlet factor added random error. Examining the reliability estimates based on the mean standard error, the bi-factor, testlet-effects, and polytomous models accurately estimated this decrease in reliability while the independent-items model inflated the reliability estimate. For the real data, the reliability estimates were higher when the scores were estimated with the independent-items model, but given the simulation results it is reasonable to assume these reliability estimates were spuriously high.

For the simulated data, the RMSE was not consistently higher for any estimation model in the middle ranges of ability. At the lower end of the trait range, the testlet-as-polytomous-item model had the larger bias and RMSE, likely due to the loss of information when the items were summed within each testlet. For the real data, correlations among the trait estimates from different models were very high though the RMSDs showed that the estimates for individuals were not precisely identical.

If the focus is on estimated θ's and not on item parameters, any of the models will perform satisfactorily, though the estimated reliability will be inflated for the independent-items model if items within testlets are not independent.

Item Parameter Estimates

The choice of model had a bigger impact on the item parameter estimates than on the trait estimates. If the item parameters are to be used for equating or for assembling test forms, these differences could have a practical impact. When the data were generated with the bi-factor or testlet-effects model, the item difficulties were recovered well, though with larger RMSE, using the independent-items model, but the item slopes were biased, consistent with previous findings of others (Ackerman, 1987; Bradlow et al., 1999; Wainer et al., 2000; Wainer & Wang, 2000). When the independent-items model was used for estimation, the RMSE of the estimates of the difficulty and the discrimination increased as the testlet slope increased. The size of the testlet slope had little impact on the accuracy of the bi-factor estimates. When the data followed the bi-factor model, the RMSE for the testlet-effects model were generally larger for the slopes but not for the difficulties. When the data followed the testlet-effects model, the RMSE for the bi-factor model was about the same as the RMSE for testlet-effects model.

When the data were generated with the independent-items model, adding the extra parameters of the bi-factor where they were not needed over-capitalized on chance and increased the error variance. The same was not true of the testlet-effects model, possibly because it introduced fewer parameters or because the proportionality constraints prevented the isolated testlet slopes from getting very large. To avoid the larger error introduced by using the bi-factor model when the items within a testlet are independent, Yen's Q_3 (Yen, 1984) could be used to check for the violation of local independence within each testlet, and the bi-factor model could be used only for selected testlets that showed larger dependencies.

In general, using the most complex model when the least complex model was adequate led to slightly higher RMSE but not to bias. Using the least complex model when either of the more complex models was appropriate led to an increased RMSE and to negatively biased slopes. Using the middle-complexity testlet-effects model when either the more or less complex model was appropriate led to only small increases in RMSE. Generally, this would be evidence favoring the use of the more parsimonious testlet-effects model over the bi-factor model. However, given the ease and speed with which the bi-factor model can be run in commercial software, in contrast to the testlet-effects model, applied practitioners may prefer to use the more complex bi-factor model. The additional parameters in the bi-factor model do not appear to decrease the accuracy of the primary trait or slope estimates, even when the data follow the more constrained testlet-effects model. These additional parameters do increase the RMSE when the items are independent, but the increase is small and could perhaps be avoided by specifying a testlet slope only for selected testlets where it is most needed.

Note

[1] The use of the χ^2 difference test may not be strictly appropriate in this context because, even though the models are nested, the item parameters are estimated by different procedures. In trial runs where the bi-factor and unidimensional models were estimated through MCMC procedures the parameter estimates were quite close to those obtained using MML estimation, so this was not a substantial issue with these data.

References

Ackerman, T. A. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED284902).

Adams, R., & Wu, M. (Eds.). (2002). *PISA 2000 technical report*. Paris: Organization for Economic Cooperation and Development. Available from http://pisaweb.acer.edu.au/oecd/oecd_pisa_data.html.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

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

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

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* [Computer manual]. Lincolnwood, IL: Scientific Software International.

Dresher, A. R. (2004). *An empirical investigation of LID using the testlet model: A further look*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika, 57,* 423–436.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271–287). Dordrecht, Netherlands: Kluwer.

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

Li, Y., Bolt, D. M., & Fu, J. (2004). *A comparison of alternative models for testlets*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum Associates.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.

Muraki, E., & Bock, R. D. (2003). *PARSCALE 4.1* [Computer software and manual]. Lincolnwood, IL: Scientific Software International.

Organization for Economic Cooperation and Development (2002). *Programme for International Student Assessment 2000* [Data file]. Available from http://pisaweb.acer.edu.au/oecd/oecd_pisa_data.html.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S\text{-}\chi^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27,* 289–298.

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

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS 1.4 [computer software]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37,* 58–75.

Stone, C. A. (2003). Empirical power and Type I error rates for an IRT fit statistic that considers the precision of ability estimates. *Educational and Psychological Measurement, 63,* 566–583.

Stone, C. A., & Hansen, M. A. (2000). The effect of errors in estimating ability on goodness-of-fit tests for IRT models. *Educational and Psychological Measurement, 60,* 974–991.

Stout, W., Douglas, J., Junker, B., & Roussos, L. (1999). *DIMTEST* [Computer software and manual]. Champaign, IL: The William Stout Institute for Measurement.

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

Wainer, H. (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,* 157–186.

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

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

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

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29,* 296–318.

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,* 125–146.

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

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39,* 291–309.

Author

CHRISTINE E. DEMARS is an Associate Professor of Graduate Psychology/Associate Assessment Specialist, Center for Assessment and Research, James Madison University, MSC 6806, Harrisonburg, VA 22807; demarsce@jmu.edu. Her primary research interests include applications of item response theory.