Confirming Testlet Effects

Christine E. DeMars

Applied Psychological Measurement 2012 36: 104
DOI: 10.1177/0146621612437403

Article

Confirming Testlet Effects

Applied Psychological Measurement
36(2) 104–121

DOI: 10.1177/0146621612437403
http://apm.sagepub.com
SAGE

Christine E. DeMars[1]

Abstract

A testlet is a cluster of items that share a common passage, scenario, or other context. These items might measure something in common beyond the trait measured by the test as a whole; if so, the model for the item responses should allow for this testlet trait. But modeling testlet effects that are negligible makes the model unnecessarily complicated and risks capitalization on chance, increasing the error in parameter estimates. Checking each testlet to see if the items within the testlet share something beyond the primary trait could therefore be useful. This study included (a) comparison between a model with no testlets and a model with testlet *g*, (b) comparison between a model with all suspected testlets and a model with all suspected testlets *except* testlet *g*, and (c) a test of essential unidimensionality. Overall, Comparison b was most useful for detecting testlet effects. Model comparisons based on information criteria, specifically the sample-size adjusted Bayesian Information Criteria (SSA-BIC) and BIC, resulted in fewer false alarms than statistical significance tests. The test of essential unidimensionality had true hit rates and false alarm rates similar to the SSA-BIC when the testlet effect was zero for all testlets except the studied testlet. But the presence of additional testlet effects in the partitioning test led to higher false alarm rates for the test of essential unidimensionality.

Keywords

testlets, bifactor, dimensionality

Detecting Testlet Effects

Test items are sometimes grouped into sets, or testlets, that share a common scenario or reading passage. This saves testing time, as the examinees only have to consider the scenario once and then can use information from it for several items. This might also increase the authenticity of the task as it adds more context. However, it could potentially introduce additional sources of variance, additional factors due to the testlets. These testlet factors would be considered random nuisance factors and not be of interest in themselves. However, ignoring the testlet factors could lead to incorrect estimation of the reliability or standard error of the primary trait measured by the test as a whole (Bradlow, Wainer, & Wang, 1999; Marais & Andrich, 2008a; Sireci, Thissen, & Wainer, 1991; Wainer & Wang, 2000; Yen, 1993). In addition, ignoring the testlet factors can lead to errors in equating/scaling (Bishop & Omar, 2002; Lee, Kolen, Frisbie, & Ankenmann, 2002; Y. Li, Bolt, & Fu, 2005), misestimation of item discrimination parameters (Ackerman, 1987; Bradlow et al., 1999; Wainer & Wang, 2000), and item misfit

[1]James Madison University, Harrisonburg, VA, USA

Corresponding author:
Christine E. DeMars, Center for Assessment and Research Studies, MSC 6806, James Madison University, Harrisonburg, VA 22807, USA
Email: demarsce@jmu.edu

(Marais & Andrich, 2008a). Even if the item parameter estimates from a model ignoring the testlets were the same as those from a model incorporating the testlet factors, the information function from the unidimensional model would be inaccurate because it should be averaged over the distribution of the testlet factor (Ip, 2010b; Wainer & Wang, 2000).

Two general conceptions may be used to model the testlet factors: local item dependency and multidimensionality. Defining local dependence in terms of nonzero covariances among the observed item responses after controlling the primary trait, Ip (2010a), showed that multidimensional and locally dependent unidimensional models yield identical covariance matrices for the observed (manifest) item responses. Local dependence can also be defined in terms of dependence of the response function for one item on the *observed* response to another item (Andrich & Kreiner, 2010; Marais & Andrich, 2008a, 2008b). This type of local dependency cannot be captured by a multidimensional item response theory (IRT) model. Dependency on the observed response might be most likely when the response to one item depends directly on the correctness of another item; for example, in a testlet linked to a graph of the results of a lab experiment, one item might ask examinees *what* occurred in the graph, followed by an item asking *why* it occurred. Multidimensionality might seem more reasonable when there are multiple items linked to a context but they do not build directly off of each other.

From the multidimensional perspective, the bifactor model or a random-effects model (Bradlow et al., 1999) can be used to model the testlet factors, as nuisance parameters, along with the primary trait. However, if one of the testlets does not measure a secondary factor, the bifactor model will merely capitalize on chance for that testlet, producing larger standard errors of the estimated item parameters (DeMars, 2006). Content specialists may also waste time speculating on why some items within the testlet loaded more highly than did others on the nonexistent secondary factor. It would therefore be useful to have a method for testing each testlet to see if it significantly measures a secondary factor. This study examines two possible sets of procedures, one based on model fit in Testfact (Bock et al., 2003) and the other based on a confirmatory test in Dimtest (Stout, 2005).

Bifactor Model

In the bifactor model (Gibbons & Hedeker, 1992), each item response is a function of the primary trait θ_P and possibly one testlet trait,[1] θ_{Tg}, where g is the testlet to which the item belongs. To identify the model, typically the mean and variance are set to 0 and 1 for each trait. All traits or dimensions in the model are orthogonal and thus θ_{Tg} is some variance that the items within testlet g share above and beyond the primary factor. In addition, responses to items that are independent of any testlets can be modeled as a function of only θ_P. For dichotomous testlet items, the 3-parameter bifactor model is

$$\Pr_i(\boldsymbol{\theta}) = c_i + (1 - c_i)\Phi\left(a_{Pi}\theta_P + \mathbf{a}'_{Ti}\boldsymbol{\theta}_T - d_i\right), \quad (1)$$

where $\Pr_i(\boldsymbol{\theta})$ is the probability of correct response on item i given the item parameters and $\boldsymbol{\theta}$ (composed of θ_P and the vector θ_T of the testlet traits), c_i is the lower asymptote, a_{Pi} is the item discrimination on the primary trait, $\mathbf{a_{Ti}}$ is a vector of testlet discrimination parameters for item i, and d_i is the item difficulty. The symbol Φ indicates integration over the normal curve up to $a_{Pi}\theta_P + \mathbf{a}'_{Ti}\boldsymbol{\theta}_T - d_i$. For any item i, all but one of the testlet a's is equal to zero, and thus only one element of θ_T has an impact on the function. For brevity, a_T will refer to the nonzero element of $\mathbf{a_{Ti}}$ (for item i within testlet g, this would more formally be $\mathbf{a}_{T(g)}$). Similarly, θ_T will indicate the element of $\boldsymbol{\theta}_T$ that is relevant for a given testlet. The subscript i will also be dropped from a_{Pi} and d_i for readability.

The random-effects testlet model (Bradlow et al., 1999; Wainer, Bradlow, & Wang, 2007) can be shown to be a constrained reparameterized version of the bifactor model (DeMars, 2006; Li, Bolt, & Fu, 2006).This model is typically parameterized as

$$\Pr_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i + \gamma_{g(i)})}}{1 + e^{a_i(\theta - b_i + \gamma_{g(i)})}}. \quad (2)$$

Essentially, in the random-effects testlet model, a_{Pi} is outside of the parentheses and thus applies to the testlet trait as well as the primary trait and thus no longer needs the subscript *P*. In addition, a_i will be approximately 1.7 times a_{Pi} because Equation 1 is in the normal metric and Equation 2 is in the logistic metric. The item difficulty, b_i, equals d_i/a_i from Equation 1. The testlet trait for testlet *g*, of which item *i* is a member, is symbolized as $\gamma_{g(i)}$ (instead of θ_T), and its coefficient is fixed to one so that the variance of γ is a free parameter. The product of the a_P and the standard deviation of γ thus equals a_T in the bifactor model. A larger variance of γ, corresponding to a larger testlet *a*-parameter in the bifactor model, indicates a greater testlet effect. The key difference from the bifactor model is that all items in the same testlet have the same a_P/a_T ratio, equal to the standard deviation of γ.

The random-effects testlet model is also equivalent to a higher order model with one higher order factor and a second-order factor for each testlet (Rijmen, 2010). The a_i of the random-effects testlet model is equivalent to the product of the loading of item *i* on the testlet second-order factor times the loading for the testlet second-order factor on the first-order factor. The standard deviation of γ in the random-effects testlet model is equivalent to the loading of item *i* on the testlet second-order factor.

Testfact and the −2 Log-Likelihood (−2LL) Difference Test

Testfact (Bock et al., 2003) can be used to estimate the parameters of the bifactor model. Thus, it seems reasonable, when using Testfact to estimate the parameters, to use Testfact to test model fit for each testlet. The estimation method in Testfact is marginal maximum likelihood (MML). After the final iteration of the item parameter estimation, the LL of each response pattern, given the parameter estimates, is calculated as a function of θ. The model LL is the sum across examinees. The better the model fit, the higher the LL or the lower the −2LL.

When comparing nested models, where one model is a more constrained version of the other model, the difference in the −2LL should be distributed as χ^2 with degrees of freedom equal to the number of additional free parameters. This test is also labeled the likelihood ratio test (LRT) because the difference in the LLs is the log of the ratio of the likelihoods. When a unidimensional model is compared with a model with one testlet, the number of parameters increases by the number of items in testlet *g*. This test is illustrated for the bifactor model, testing multiple secondary factors simultaneously against a unidimensional model, in the Testfact manual (duToit, 2003) and other sources (Bock & Gibbons, 2010; Cai, 2010; Gibbons & Hedeker, 1992). Two possible model comparisons could be conducted to test an individual testlet. If one ignores the possibility of testlet traits for testlets other than testlet *g*, a bifactor model could be compared with a unidimensional model. In the unidimensional model, all testlet discriminations are fixed to zero, and in the bifactor model considered here, labeled the *single-testlet* model, the testlet discriminations for testlet *g* are free. Thus, the unidimensional model is nested within the single-testlet bifactor model. A rejection of the null hypothesis supports the presence of a significant testlet factor for testlet *g*. An alternative model comparison, between a *complete* bifactor model and an *all-but-one* bifactor model, could take into account that other testlets may be present. In the complete bifactor model, testlet discriminations are freed for each possible testlet. In

the all-but-one model the testlet discriminations for testlet *g* are fixed to zero and the items within testlet *g* load only on the primary trait. The all-but-one model is nested within the complete model, so again the difference in the −2LL should be distributed as χ^2 with degrees of freedom equal to the number of additional freed parameters (the number of items in testlet *g*). The comparison between the unidimensional and the single-testlet model is appropriate if the null hypothesis is completely true or if testlet *g* is the only testlet significantly influenced by a testlet trait. This comparison avoids capitalizing on chance and wasting degrees of freedom on the other testlets in which there is no true testlet effect. However, it might be problematic in cases where there are significant testlet effects for testlets other than testlet *g*—the comparison of the complete versus the all-but-one models would then be more appropriate.

When Testfact is used for exploratory models, in which the second factor is not based on testlet structure or any other theoretical reason but is instead chosen mathematically to explain the greatest amount of residual variance, the difference in −2LL sometimes has been shown to have an extremely high Type I error rate (De Champlain & Gessaroli, 1998; DeMars, 2003), particularly for less discriminating items (Berger & Knol, 1990), although this finding has not been universal (Tate, 2003). Hayashi, Bentler, and Yuan (2007), in the context of direct maximum likelihood estimation of linear models, explained that this was due to rank deficiency in the more complex model because one column of factor loadings is actually zero. This causes the distribution of the −2LL difference to be shifted to the right of the assumed χ^2. Rank deficiency should also be a problem for the bifactor model when the true loadings on the testlet factor are all zero. Nevertheless, because the difference test is recommended for the bifactor model in the software guide and other recent sources (Bock & Gibbons, 2010; Cai, 2010; Gibbons & Hedeker, 1992), it was included in this study to assess how badly the difference in −2LL might depart from the χ^2 distribution.

Testfact and Information Criteria for Model Selection

In addition to statistical significance testing, the models can be compared based on information criteria, including Akaike's Information Criterion (AIC; Akaike, 1987), Bayesian Information Criterion (BIC; Schwarz, 1978), and the sample-size adjusted BIC (SSA-BIC; Sclove, 1987). Instead of testing whether the more complex model fits *significantly* better, these indices compare which model fits better, after a penalty for the number of parameters (*p*) estimated, with no statistical significance test:

$$\mathrm{AIC} = -2\mathrm{LL} + 2p. \tag{3}$$

The AIC is not asymptotically consistent (Bozdogan, 1987; Woodroofe, 1982); it does not take the sample size (*N*) into account and thus does *not* become increasingly accurate with sample size. Instead, it tends to favor more complex models with larger sample sizes. To address this issue, with the BIC and SSA-BIC, the penalty for model complexity increases for large samples:

$$\mathrm{BIC} = -2\mathrm{LL} + p(\ln(N)). \tag{4}$$

The BIC is asymptotically consistent (Haughton, 1988; Schwarz, 1978; Woodroofe, 1982). In the SSA-BIC, *N* is replaced with (*N* + 2)/24, so the sample-size adjustment is less severe. The AIC, BIC, and SSA-BIC are transformations of the −2LL, so lower values indicate better fit. When using these model comparison indices, the model with the lowest value is selected—there is no statistical significance test. These indices are not included in the Testfact output but can easily be calculated from the −2LL.

The information-criterion indices can be used to compare any models estimated by maximum likelihood. One study specifically used information-based indices with the bifactor model: Y. Li and Rupp (2011) found that the mean AIC and BIC were both lower (better) for the bifactor model than the unidimensional model when the data were generated from a bifactor model. They did not report the percentage of data sets for which the AIC and BIC were lower.

Several other researchers have applied the AIC and BIC in other IRT contexts, although not specifically for the bifactor model. McKinley (1989) studied the AIC and the consistent AIC (CAIC, similar to the BIC) for selecting multidimensional models. For a unidimensional data set, the AIC was lower for a more complex model, and for a three-dimensional data set both indices chose the correct model. Kang and Cohen (2007) compared 1 parameter logistic (1PL), 2PL, and 3PL unidimensional models. Both AIC and BIC were effective when the true model was the 1PL or 2PL. But when data followed a 3PL model, the BIC always erroneously preferred the 1PL or 2PL, and the AIC also tended to choose the 2PL when the test was easy and had few examinees near the lower asymptote. Kang, Cohen, and Sung (2009) used the AIC and BIC to compare polytomous IRT models; the BIC was more accurate. F. Li, Cohen, Kim, and Cho (2009) compared the AIC and BIC, as well as several Bayesian indices, for selecting IRT mixture models. BIC was generally most accurate; as sample size increased, AIC tended to favor more complex models.

The information-criterion indices have also been applied with real data to choose among IRT models. McKinley and Way (1992) used the AIC and the CAIC to compare multidimensional models for data from an English language test. Janssen and De Boeck (1999) compared multidimensional IRT models for tests of synonyms with the AIC and CAIC. Semmes, Davison, and Close (2011) used the AIC and BIC to compare a model with a random effect for ability to a model with an additional fixed effect for speed and a model with random effects for both ability and speed.

The SSA-BIC has not seen widespread use in IRT, but it appears to be reasonably effective in selecting the number of latent classes in mixture modeling or latent class analysis (Nylund, Asparouhov, & Muthén, 2007; Yang, 2006; Yang & Yang, 2007). The SSA-BIC penalty for estimating additional parameters is greater than the AIC penalty but less than the BIC penalty. Thus, the SSA-BIC is less likely than the AIC, but more likely than the BIC, to choose the more complex model with increasing sample size. Yang (2006) found the BIC and SSA-BIC were comparable at sample sizes of 1,000, but the SSA-BIC was more accurate at sample sizes of 500, where the BIC tended to choose the model with fewer parameters. In Nylund et al., the BIC and SSA-BIC were very accurate with samples of 500 or 1,000, but with samples of 200 the SSA-BIC showed a tendency to erroneously pick more complex models, although to a lesser extent than the AIC.

Dimtest and Essential Unidimensionality

Bifactor models are not the only approach for modeling testlets. Sometimes polytomous IRT models are used (Bishop & Omar, 2002; Marais & Andrich, 2008a; Sireci et al., 1991; Zenisky, Hambleton, & Sireci, 2002), with all items in the testlet treated as a single item. Either the items are summed so an ordinal model may be used or each response pattern within the testlet is treated as a separate item response for a nominal model. This approach is more consistent with the conceptualization of the correlations as local dependence. Yet another approach is to use observed number-correct scores and adjust the reliability estimate for the dependency (Sireci et al., 1991). If one plans to use either polytomous IRT or number-correct scoring, the testlet effects could be assessed by a test for essential unidimensionality. Jang and Roussos (2007) used this approach to check each of the passage-based testlets on a reading comprehension test.

Stout (1987) used the term *essential unidimensionality* to indicate zero mean covariances among the responses, conditional on the primary trait. This is a necessary, but not sufficient, condition for strict conditional independence. Dimtest (Stout, 2005) tests whether the average interitem covariance for items within the assessment test differs from zero after controlling for scores on the partitioning test. The assessment test is the cluster of items suspected of measuring a secondary dimension; in this context, a testlet would comprise an assessment test. The partitioning test is either the remaining items or a designated group of items believed to be unidimensional. Calculation of Stout's T begins by segmenting the examinees by their score on all items except the items in testlet g. Within each score group k, $T_{L,k} = \hat{\sigma}_k^2 - \hat{\sigma}_{U,k}^2$, where $T_{L,k}$ is the sum of the interitem covariances for testlet g within score group k, calculated as the difference between $\hat{\sigma}_k^2$, the variance of the testlet g scores, and $\hat{\sigma}_{U,k}^2$, the sum of the item variances for group k for the items within testlet g. These covariance sums are averaged across the K score groups: $T_L = 1/\sqrt{K} \sum_{k=1}^{K} T_{L,k}/S_k$, where S_k is the standard deviation of $T_{L,k}$ (see Stout, Froelich, & Gao, 2001, for the calculation of the standard deviation). For short tests, T_L is positively biased. A bias correction, $\bar{T}_G$, is calculated using simulation methods (Stout et al., 2001). The test statistic is $T = T_L - \bar{T}_G / \sqrt{1 + 1/N}$, where N is the number of examinees. After this bias correction, Stout's T approximates a Z distribution when the null hypothesis is true.

Other Methods for Assessing Testlet Dependency

Testlets have also been approached by examining the dependence of pairs of items within the testlet. Yen (1984, 1993) proposed Q_3 as an index of whether a pair of items was locally dependent. Q_3 is simply the linear correlation between the residuals for item i and item j. Andrich and Kreiner (2010) developed another pairwise index of local dependence: the difference in the difficulty parameter of item j for those who answered item i correctly compared with those who answered item i incorrectly. Alternative pairwise indices have also been described in Chen and Thissen (1997); in Kim, De Ayala, Ferdous, and Nering (2011); and in Rosenbaum (1988). Pairwise indices will not be considered further here; instead, the focus is on the testlet level for testlets containing more than two items.

Method

Data Simulation

Three test forms were simulated: a 25-item test with 5 items within each of 5 testlets, a 50-item test, with 5 items within each of 10 testlets, and a 50-item test with 10 items within each of 5 testlets. The two 50-item conditions allowed the exploration of increasing the number of testlets compared with increasing the number of items within testlets. Item responses followed the bifactor model as specified in Equation 1. Different item parameters were randomly selected for each replication. The d-parameters were randomly selected from a normal distribution with mean equal to 0 and standard deviation equal to 1, restricted to the range -3 to 3. All c-parameters were given a value of 0.2. The natural logs of the a_P-parameters were randomly selected from a normal distribution with mean equal to 0 and standard deviation equal to 0.5, with the resulting a_P-parameters restricted to the range 0.5 to 2.0.

The ratio a_T/a_P was set to a constant within each testlet, to make it simple to quantify the extent of the testlet effect. This made the bifactor model equivalent to the random-effects testlet model (standard deviation of $\gamma = a_T/a_P$, $b = d/a_P$). The ratios were (0.0, 0.3, 0.6, 0.9, 1.2). When

there were 10 testlets, 2 testlets had the same ratio, for convenience in summarizing the results, but they were related to independent θ_T. These ratios were within the ranges reported in the literature. For example, on an analytical reasoning test, Y. Li et al. (2006) reported γ variances of 1.27, 0.75, 0.82, and 2.10, equivalent to a_T/a_P = 1.13, 0.87, 0.91, and 1.45. On two verbal tests, Wainer, Bradlow, and Du (2000) reported estimated γ variances ranging from 0.11 to 0.96, corresponding to a_T/a_P ranging from 0.33 to 0.98. A ratio of 0 indicates that the items within the testlet share no variance beyond that attributable to the primary factor, and a ratio greater than 1 indicates that the testlet trait has more influence on the response than the primary trait.

First, to assess the Type I error rate for the significance tests and the false alarm rate for the information-criterion indices, all a_T were set to zero for all 5 or 10 testlets. Next, in one set of the studies, only one of the 5 or 10 testlets at a time had nonzero a_T. For one set of replications, a_T/a_P = 0.3 for one testlet while a_T/a_P = 0 for the other testlets, followed by another set of replications where a_T/a_P = 0.6 for one testlet while a_T/a_P = 0 for the other testlets, and so on. This methodology is analogous to testing an item for differential item functioning (DIF) when all other items are DIF free. Finally, in another set of studies, each testlet had a different a_T/a_P at the same time. This more realistic set allowed for testing each testlet effect in the presence of effects in other testlets.

Data for 2,000 examinees was simulated 1,000 times for each of the conditions. For each simulee, θ_P and each element of the vector θ_T were drawn from independent distributions, $N(0, 1)$. Probability of correct response was calculated from θ_P, θ_T, a_T, a_P, and d, and the simulated response was coded correct if the probability was greater than a random draw from a uniform (0, 1) distribution.

Testfact and −2LL

Two pairs of models were run for each of the data sets: the unidimensional model versus the single-testlet bifactor model and the all-but-one bifactor model versus the complete bifactor model.

Default prior distributions and starting values were used for the a_P, a_T, and d-parameters. Although the data were simulated with a constant within-testlet a_T/a_P ratio, the ratio was not fixed in the estimation. In real data, a constant a_T/a_P ratio may not fit well (Y. Li et al., 2006; Rijmen, 2010); it was only fixed in the generation of the data so that the extent of the testlet effect could systematically vary over testlets. The c-parameters were fixed to the true value of 0.2. The maximum number of expectation-maximization (EM) cycles was increased to 100, with a stopping criterion of .005. Nine quadrature points were used for each dimension (81 points in total because no item loaded on more than two dimensions).

Statistical significance was tested with the difference between the models in −2LL, at α = .01. The null hypothesis was as follows:

$$H_0: -2LL_{reduced} - (-2LL_{full}) = 0. \quad (5)$$

In the single-testlet conditions, the reduced model was the unidimensional model and the full model was a bifactor model with all items loading on the primary factor and the items for the studied testlet additionally loading on a secondary factor. In the all-but-one conditions, the reduced model was a bifactor model with a secondary factor for each testlet *except* the studied testlet, whose items loaded only on the primary factor. The full model included secondary factors for every testlet. The difference was compared to a χ^2 distribution, with degrees of freedom equal to 5 or 10, the testlet length.

In addition, the same full and reduced models were compared using the AIC, BIC, and SSA-BIC, calculated from the −2LL. For each index, the model with the lower value had better fit. These model comparisons do not involve statistical significance testing.

Dimtest's Test of Essential Unidimensionality

As in Testfact, each testlet was tested separately in Dimtest. The studied testlet was used as the assessment test and all other items comprised the partitioning test, with a minimum cell size of two. The true *c* of 0.2 was specified. For the bias correction, 50 kernel-smoothing points and 100 replications were used. Stout's *T* statistic, as printed in the output, was used to compute Type I error or power, at $\alpha = .01$. A total of 5 or 10 comparisons, one for each testlet, were conducted for each replication of each condition.

Conceptually, the null hypothesis corresponding to $T = 0$ is

$$H_0 : \Sigma\sigma_{ij}|PT = 0, \tag{6}$$

where *PT* is the true score on the partitioning test and the mean conditional covariance is taken over all items *i* and *j* within testlet *g*, and integrated (averaged) over all values of *PT*. *T* is approximately normally distributed under the null hypothesis.

Results

False Alarm Rates When No Testlet Had Nonzero a*-Parameters*

The hit rate for this study was defined as the proportion of times that the more complex model was selected. This selection was through the statistical significance test for the −2LL difference and for the Dimtest statistic. For the AIC, SSA-BIC, and BIC, the model with the lowest value was selected. The terms *Type I error* and *power* do not strictly apply to these latter comparisons; the more generic terms *false alarm rate* and *true hit rate* would be more correct. However, for clarity, the term *power* is sometimes used here for the true hit rate. When the a_T/a_P ratio was 0, selection of the more complex model was termed a false alarm (false positive, false hit). For the other a_T/a_P ratios, this was termed a true hit.

First, the false alarm rates or Type I error rates were calculated for conditions where the data were unidimensional. Table 1 shows these false alarm rates. In the case of the test of the difference in −2LL and Stout's *T* in Dimtest, which are statistical significance tests, this is the empirical Type I error rate. The nominal Type I error rate was .01. The Testfact difference in −2LL test was liberal, and was similar for the two pairs of models (unidimensional vs. one testlet, all-but-one vs. complete). It became more liberal when the test length increased, and it was worse with 5 testlets of 10 items than with 10 testlets of 5 items. There was more capitalization on chance for longer testlets. For Dimtest, the Type I error rate was lower than the nominal rate for all three test forms. The AIC erroneously accepted the more complex model far too frequently, following the same pattern as the difference in −2LL test. Use of the BIC and SSA-BIC seldom led to false selection of the more complex model.

The −2LL difference test is purported to follow a χ^2 distribution, and the Dimtest *T* is tested against a *Z* distribution. The values in Table 1 only provide information about the tail of the test statistic distribution. Figure 1 compares the empirical distribution of the −2LL difference test to a χ^2 distribution with 5 or 10 degrees of freedom as appropriate. When there were 25 items in 5 testlets, the −2LL was a bit less skewed than the χ^2 distribution, pulling more of the distribution into the rejection range in the right tail. This effect became somewhat worse when the

Table 1. False Alarm Rate When No Testlet Has Nonzero a_T / a_P

	Index								
	Testfact								
	Unidimensional–single testlet				All-but-one–complete				Dimtest
	−2LL diff[a]	AIC	BIC	SSA-BIC	−2LL diff[a]	AIC	BIC	SSA-BIC	T^a
25 items: 5 testlets	.024	.142	.000	.000	.027	.150	.000	.001	.007
50 items: 10 testlets (5 items per testlet)	.049	.220	.001	.006	.036	.174	.000	.003	.003
50 items: 5 testlets (10 items per testlet)	.210	.347	.000	.001	.173	.295	.000	.000	.000

Note: LL = log likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = sample-size adjusted BIC.

[a]The false alarm rate is the Type I error rate for the −2LL difference test and for Stout's *T*, with nominal $\alpha = .01$.

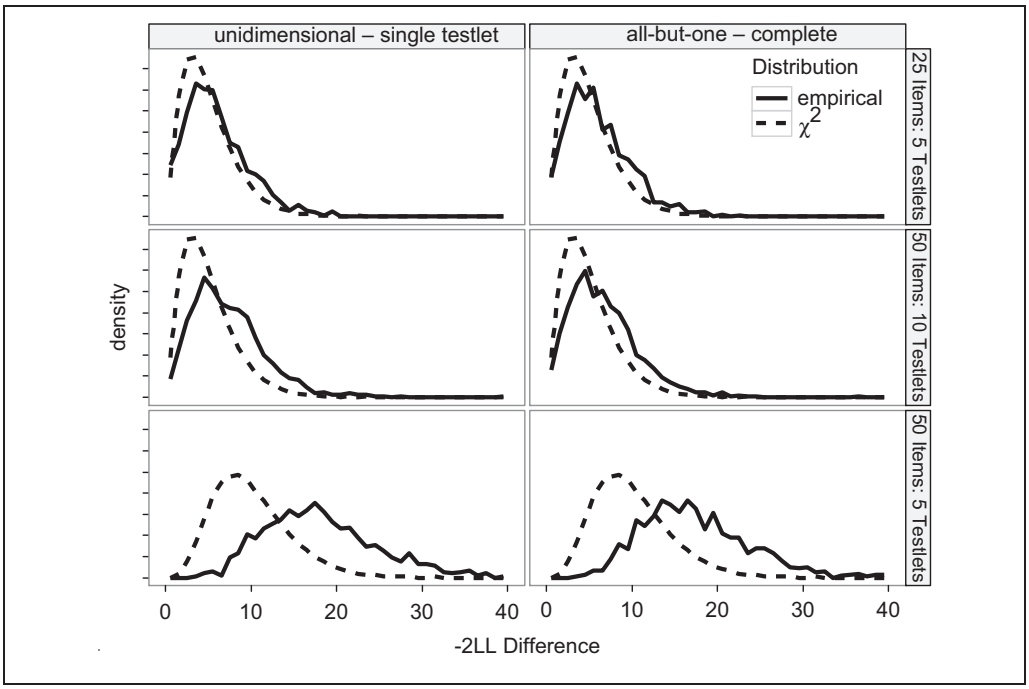

Figure 1. Distribution of the difference in −2 log likelihood (−2LL) when $a_T = 0$ for all items in all testlets

Note: In the left panels, a unidimensional model is compared with a model with a_T for items in one testlet. In the right panels, a model with a_T for all items except those in the studied testlet is compared with a model with a_T for all items.

test length was increased by adding more testlets (50 items, 10 testlets), but it became much worse when the test length was increased by adding more items per testlet (50 items, 5 testlets). For longer testlets with 10 items, the distribution did not remotely follow a χ^2 distribution with 10 degrees of freedom.

The left panel of Figure 2 shows comparable information for Dimtest. Stout's *T* was slightly more leptokurtic than *Z*, so fewer replications fell in the rejection range. With the 25-item, five-testlet form, however, this was counteracted because the empirical distribution was slightly off-center, providing a good match in the right tail.

Power Rates When Only One Testlet Per Form Had Nonzero a-Parameters

Table 2 shows the hit rates, or power rates, for the test forms in which all testlets except the studied testlet had testlet *a*-parameters of zero. The studied testlet had an a_T/a_P ratio ranging from 0.3 (a small testlet effect) to 1.2 (true testlet discriminations were higher than true primary discriminations—a large testlet effect). Within each test form, a_T was 0 for all testlets except the studied testlet, with different data sets generated for studying each testlet in turn. All cells in the conditions with five testlets were based on 1,000 replications; cells in the conditions with 10 testlets were based on 2,000 replications because results were combined for the testlets with the same a_T/a_P ratio.

Hit rates should increase as the a_T/a_P ratio increases. Power was very high for all indices when the a_T/a_P ratio was 0.6 or higher. For the smallest a_T/a_P ratio of 0.3, the AIC was the most

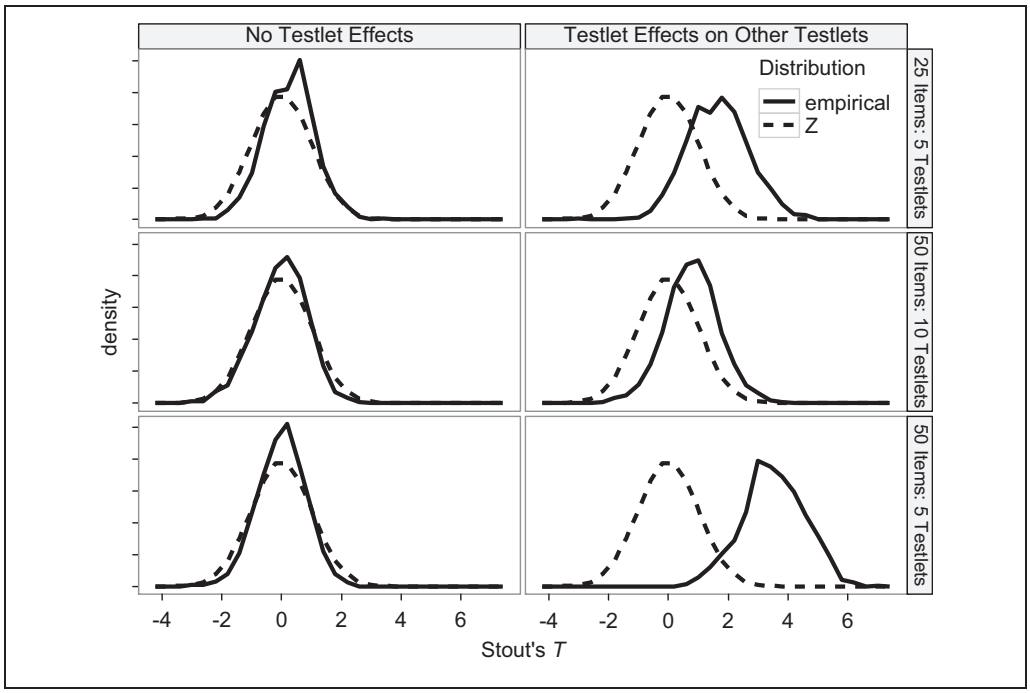

Figure 2. Distribution of the test statistic for Dimtest
Note: In the left panel, none of the testlets have $a_T/a_P > 0$. In the right panel, the studied testlet has $a_T/a_P = 0$, but the other testlets have $a_T/a_P > 0$.

powerful but it should not be trusted due to its high false alarm rate. The −2LL difference test had the next highest power, but again it was too liberal in the null condition, especially with longer testlets, so the power rate is misleading. SSA-BIC and Dimtest had roughly comparable power for the ratio of 0.3, much higher than BIC.

False Alarm Rates When Other Testlets Had Nonzero a*-Parameters*

In Table 3, $a_T = 0$ for the studied testlet, but it ranged from 0.3 to 1.2 for the other testlets in the same data set. This is somewhat like testing for DIF when the total score is contaminated by other DIF items. For the test of the all-but-one model versus the complete model, the false alarm rates for the −2LL difference test and the information criteria were similar to the false alarm rates in Table 1. But the false alarm rates for the unidimensional model versus the single testlet model, and for Dimtest, were much larger than those in Table 1. Figure 3 shows the empirical distribution of the −2LL difference test for the testlets with $a_T = 0$. In the left panel, the distributions for the unidimensional model versus the single testlet model are clearly quite far from the χ^2 distribution. In the right panel, the distribution for the all-but-one-testlet versus complete testlet model is similar to Figure 1. The right panel of Figure 2 shows the empirical distribution of Stout's T for the testlets with $a_T = 0$. The distribution is pulled to the right, and the effect is worse with fewer but longer testlets.

The problem with all of the model comparisons of the unidimensional model to a single testlet is that, if some of the other testlets have nonzero a_T, the θ measured by the unidimensional model is not the θ_P used to generate the data. Instead, θ from the unidimensional model is a

Table 2. Hit Rate (Power) When Only the Studied Testlet Has Nonzero a_T/a_P

	Index								
	Testfact								
	Unidimensional–single testlet				All-but-one–complete				Dimtest
a_T/a_P for studied testlet	−2LL diff[a]	AIC	BIC	SSA-BIC	−2LL diff[a]	AIC	BIC	SSA-BIC	T^a
25 items: 5 testlets									
0.3	0.440	0.671	0.016	0.189	0.413	0.661	0.016	0.168	0.158
0.6	0.995	0.999	0.919	0.980	0.995	0.998	0.913	0.977	0.912
0.9	1.000	1.000	0.999	1.000	1.000	1.000	.999	1.000	0.997
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50 items: 10 testlets (5 items per testlet)									
0.3	0.604	0.821	0.072	0.364	0.561	0.788	0.059	0.331	0.224
0.6	0.999	1.000	0.949	0.992	0.999	1.000	0.942	0.990	0.969
0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50 items: 5 testlets (10 items per testlet)									
0.3	0.979	0.989	0.261	0.762	0.973	0.982	0.206	0.708	0.654
0.6	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: LL = log likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = sample-size adjusted BIC.

[a]Nominal α = .01 for −2LL difference test and for Stout's *T*.

Table 3. False Alarm Rate When $a_T = 0$ for the Studied Testlet but Other Testlets Have Nonzero a_T/a_P

	Index								
	Testfact								
	Unidimensional–single testlet				All-but-one–complete				Dimtest
	−2LL diff[a]	AIC	BIC	SSA-BIC	−2LL diff[a]	AIC	BIC	SSA-BIC	*T*[a]
25 items: 5 testlets	.491	.722	.050	.277	.021	.110	.000	.000	.252
50 items: 10 testlets (5 items per testlet)	.163	.421	.001	.048	.037	.150	.000	.003	.044
50 items: 5 testlets (10 items per testlet)	.993	.997	.491	.909	.136	.267	.000	.001	.871

Note: LL = log likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = sample-size adjusted BIC.
[a]Nominal $\alpha = .01$ for −2LL difference test and for Stout's *T*.

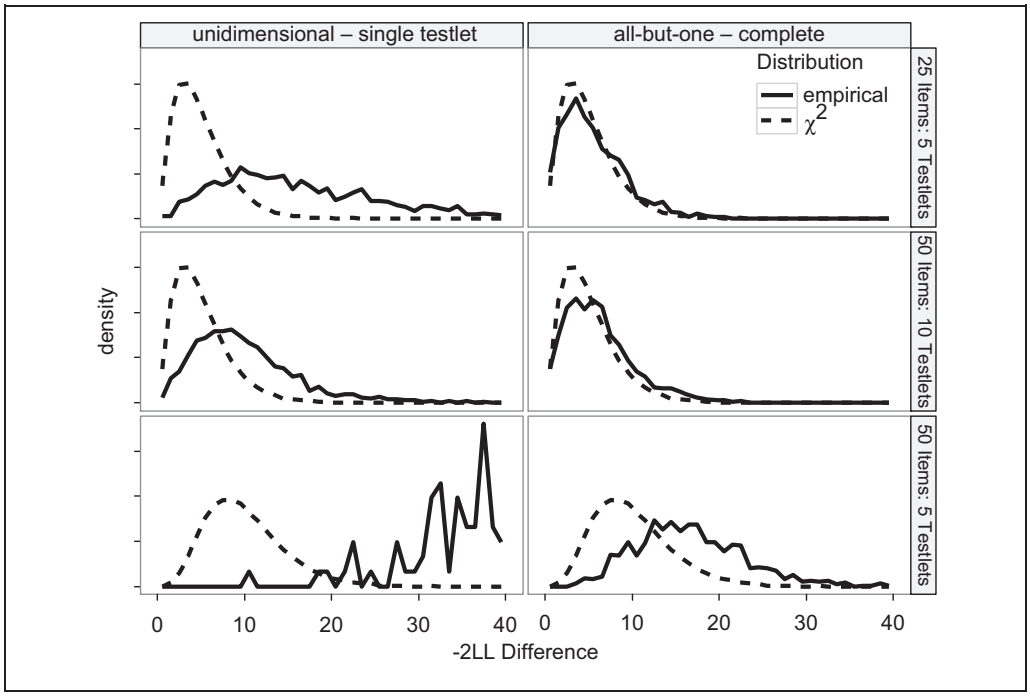

Figure 3. Distribution of the difference in −2LL when $a_T/a_P > 0$ in all testlets except the studied testlet
Note: In the left panels, a unidimensional model is compared with a model with a_T for items in one testlet. In the right panels, a model with a_T for all items except those in the studied testlet is compared with a model with a_T for all items.

composite of θ_P and the vector of θ_T. The unidimensional composite measured mostly in the direction of θ_P but was somewhat deflected toward the testlets with higher a_T/a_P ratios. This effect was stronger when each testlet contained 20% of the total test items, compared with 10%

Table 4. Hit Rate (Power) When Each Testlet Within the Test Form Has a Different a_T/a_P

	Index								
	Testfact								
	Unidimensional–single testlet				All-but-one–complete				Dimtest
a_T/a_P	−2LL diff[a]	AIC	BIC	SSA-BIC	−2LL diff[a]	AIC	BIC	SSA-BIC	T[a]
25 items: 5 testlets									
0.3	0.904	0.956	0.395	0.764	0.258	0.537	0.004	0.068	0.549
0.6	0.999	1.000	0.968	0.993	0.989	0.999	0.852	0.966	0.951
0.9	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.993
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50 items: 10 testlets (5 items per testlet)									
0.3	0.756	0.903	0.178	0.532	0.461	0.700	0.026	0.223	0.435
0.6	1.000	1.000	0.968	0.995	0.997	1.000	0.939	0.990	0.981
0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
50 items: 5 testlets (10 items per testlet)									
0.3	1.000	1.000	0.956	0.996	0.873	0.930	0.025	0.343	0.995
0.6	1.000	1.000	1.000	1.000	1.000	1.000	0.998	1.000	1.000
0.9	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note: LL = log likelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA-BIC = sample-size adjusted BIC.

[a]Nominal α = .01 for −2LL difference test and for Stout's *T*.

of the items.[2] All shared variance among the items in the testlet with $a_T = 0$ was due to θ_P, but because the unidimensional θ was not completely aligned with θ_P, the responses to the items in the testlet had a small amount of shared variance not explained by the estimated unidimensional θ. The same phenomenon occurs in Dimtest. The observed score on the partitioning test measures a composite of the primary trait and the other testlet traits, so the items measuring only the primary trait have some covariance after controlling for the partitioning test score. This could be avoided by including only one item from each suspected testlet on the partitioning test, but in this context that would mean only four items could compose the partitioning test in the five testlet conditions. The partitioning test would become a purer measure of the primary trait as either the number of testlets increased, the number of items per testlet decreased, or the maximum a_T decreased.

Power Rates When Multiple Testlets Had Nonzero a-Parameters

The true hit rate, or power, is reported in Table 4 for each testlet when tested in the presence of other testlet effects. Again, within each test form, each testlet had a different a_T/a_P ratio; power should increase as the ratio increases. However, these power rates could be spurious for any index which showed an inflated false alarm rate in Table 3. Only the BIC and SSA-BIC when used with the all-but-one model versus the complete model had acceptable false alarm rates. These indices had high power when the a_T/a_P ratio was 0.6 or higher. For $a_T/a_P = 0.3$, the power for these indices appeared to be somewhat lower than it was when only the studied testlet has nonzero a_T/a_P.

Discussion and Conclusion

Due to the complexity of the model, a relatively large sample size (N = 2,000) was used. The intent was to explore the procedures with a sample large enough to be realistic for multidimensional IRT so that model estimation difficulties would not be confounded with the accuracy of detecting the testlet trait. Further research is needed to assess performance with smaller samples that might be encountered in research or in certification testing, or with larger samples that might be available in large-scale settings. The BIC and SSA-BIC penalties for sample size might not be equally appropriate for much larger or smaller samples.

Additional testlet configurations also merit further consideration. Some possibilities include shorter testlets or testlets in which some items have zero a_T and others have high a_T independent of the a_P. The test could additionally include independent items not located in testlets and testlets of varying lengths.

Overall, the use of the SSA-BIC with the comparison of the all-but-one model versus the complete model was most accurate. For all of the indices, power or true hit rate was high for detecting testlets in which the testlet discrimination was at least moderately large relative to the primary discrimination ($a_T/a_P \geq 0.6$). Thus, failure to reject the null hypothesis, or a lower value of the information-criterion indices for the reduced model, would support a conclusion that the testlet factor was relatively small or nonexistent and the items could be treated as independent items.

In contrast, interpretation of a rejected null, or a lower value of the information-criterion indices for the more complex model, is more problematic. For the unidimensional model versus the single-testlet model in Testfact, and for Dimtest, the null may be rejected either due to a testlet factor in the testlet under consideration, or it may be due to the effects of other testlets distorting the unidimensional θ composite/partitioning test. The test of the all-but-one model versus the complete model purifies θ_P and thus is not susceptible to this problem. However, the statistical significance test of this model comparison has inflated Type I error, consistent with the explanation in Hayashi et al. (2007). The SSA-BIC for this model comparison has both a low false alarm rate and a high true hit rate. Thus, this index appears most promising. However, the SSA-BIC needs to be explored more for the bifactor model or other IRT models before definitive recommendation, as it has not been widely used outside of mixture modeling. If one prefers to err on the side of the less complex model (lower power, essentially), the BIC would be a good choice and has been more widely studied. For both the SSA-BIC and BIC, the magnitude of the penalty for model complexity may not always be appropriate in this context of rank insufficiency when all loadings on the secondary factor are zero, although they did seem to work well in this study.

The need for the study was framed in terms of the bifactor model. An alternative is to treat each testlet as a polytomous item and apply a unidimensional polytomous model (Bishop & Omar, 2002; Marais & Andrich, 2008a; Sireci et al., 1991; Zenisky et al., 2002). Typically, to use a polytomous model, the item scores are summed within the testlet. When there is no testlet effect, this summation results in a loss of information because the particular pattern of rights and wrongs within the testlet is lost (Yen, 1993). Although the loss in information from summation will generally be small compared with the spurious inflation of information when ignoring the testlet effect if it exists, it is worthwhile to first test each testlet for dependency before scoring it polytomously. The procedures described here thus could also be used as preliminary to polytomous models with items summed within testlets. Again, because the SSA-BIC for the comparison of the all-but-one model versus the complete model was powerful while controlling false alarms, this index is recommended, with the caution that it has not been widely studied except in the context of mixture modeling.

In some cases, analysts may prefer to model the testlet discriminations (or form summed items) for all testlets if any of the testlet factors are significant. Although this allows for greater capitalization on chance, it is easier to explain conceptually. Dimtest would be an appropriate choice in this situation. It has good power, and when none of the testlets have large effects, the Type I error is well controlled. Alternatively, one could use the SSA-BIC or BIC to compare a unidimensional model with a complete testlet model, instead of testing individual testlets.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. This model was not developed specifically for testlets. *Testlet trait* is used here instead of Gibbons and Hedecker's term *group factor* in keeping with the present context.
2. An angle of 0° with θ_P and 90° with each θ_T would indicate the composite measured only the primary trait. For the 25-item test and the 50-item test with 5 testlets and 10 items per testlet, the theoretical angle of the composite was 20° with θ_P, and (90°, 87°, 84°, 80°, 74°) for the θ_T vector. For the 50-item test with 10 testlets and 5 items per testlet, the angle of the composite was 14° with θ_P, and (90°, 89°, 87°, 85°, 83°, 90°, 89°, 87°, 85°, 83°) for the θ_T vector.

References

Ackerman, T. A. (1987, April). *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.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*, 181-192.

Berger, M. P. F., & Knol, D. L. (1990). *On the assessment of dimensionality in multidimensional item response theory models*. Enschede, Netherlands: University of Twente.

Bishop, N. S., & Omar, M. H. (2002, April). *Comparing vertical scales derived from dichotomous and polytomous IRT models for a test composed of testlets.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bock, R. D., & Gibbons, R. (2010). Factor analysis of categorical item responses. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 155-184). New York, NY: Routledge.

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

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

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika, 75*, 581-612.

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

De Champlain, A., & Gessaroli, M. E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education, 11*, 231-253.

DeMars, C. E. (2003). Detecting multidimensionality due to curricular differences. *Journal of Educational Measurement, 40*, 29-51.

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

duToit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

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

Haughton, D. M. A. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics, 16*, 342-355.

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling, 14*, 505-526.

Ip, E. H. (2010a). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63*, 395-416.

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

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*, 1-21.

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research, 34*, 245-268.

Kang, T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement, 31*, 331-358.

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499-518.

Kim, D., De Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2011). The comparative performance of conditional independence indices. *Applied Psychological Measurement, 35*, 447-471.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2002). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25*, 357-372.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373.

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

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

Li, Y., & Rupp, A. A. (2011). Performance of the S-χ^2 statistic for full-information bifactor models. *Educational and Psychological Measurement, 71*, 986-1005.

Marais, I. D., & Andrich, D. (2008a). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*, 105-124.

Marais, I. D., & Andrich, D. (2008b). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*, 200-215.

McKinley, R. L. (1989). *Confirmatory analysis of test structure using multidimensional item response theory (ETS-RR-89-31)*. Princeton, NJ: ETS.

McKinley, R. L., & Way, W. D. (1992). *The feasibility of modeling secondary TOEFL ability dimensions using multidimensional IRT models (ETS-RR-92–16)*. Princeton, NJ: ETS.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14*, 535-569.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*, 361-372.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53*, 349-359.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333-343.

Semmes, R., Davison, M. L., & Close, C. (2011). Modeling individual differences in numerical reasoning speed as a random effect of response time limits. *Applied Psychological Measurement, 35*, 433-446.

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

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (2005). Dimtest (Version 2.0) [Computer software]. Champaign, IL: William Stout Institute for Measurement.

Stout, W., Froelich, A., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-375). New York, NY: Springer-Verlag.

Tate, T. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement, 27*, 159-203.

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

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

Woodroofe, M. (1982). On model selection and the arc sine laws. *Annals of Statistics, 10*, 1182-1194.

Yang, C.-C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis, 50*, 1090-1104.

Yang, C.-C., & Yang, C.-C. (2007). Separating latent classes by information criteria. *Journal of Classification, 24*, 183-203.

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*, 125-145.

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

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