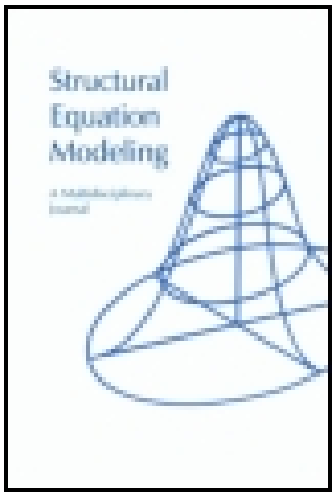

Structural Equation Modeling: A Multidisciplinary Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hsem20

When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index

Wes E. Bonifay[a], Steven P. Reise[a], Richard Scheines[b] & Rob R. Meijer[c]
[a] University of California Los Angeles
[b] Carnegie Mellon University
[c] University of Groningen
Published online: 06 Apr 2015.

To cite this article: Wes E. Bonifay, Steven P. Reise, Richard Scheines & Rob R. Meijer (2015): When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2014.938596

To link to this article: http://dx.doi.org/10.1080/10705511.2014.938596

Structural Equation Modeling: A Multidisciplinary Journal, 00: 1–13, 2015

ISSN: 1070-5511 print / 1532-8007 online
DOI: 10.1080/10705511.2014.938596

Routledge
Taylor & Francis Group

When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index

Wes E. Bonifay,[1] Steven P. Reise,[1] Richard Scheines,[2] and Rob R. Meijer[3]
[1]*University of California Los Angeles*
[2]*Carnegie Mellon University*
[3]*University of Groningen*

In structural equation modeling (SEM), researchers need to evaluate whether item response data, which are often multidimensional, can be modeled with a unidimensional measurement model without seriously biasing the parameter estimates. This issue is commonly addressed through testing the fit of a unidimensional model specification, a strategy previously determined to be problematic. As an alternative to the use of fit indexes, we considered the utility of a statistical tool that was expressly designed to assess the degree of departure from unidimensionality in a data set. Specifically, we evaluated the ability of the DETECT "essential unidimensionality" index to predict the bias in parameter estimates that results from misspecifying a unidimensional model when the data are multidimensional. We generated multidimensional data from bifactor structures that varied in general factor strength, number of group factors, and items per group factor; a unidimensional measurement model was then fit and parameter bias recorded. Although DETECT index values were generally predictive of parameter bias, in many cases, the degree of bias was small even though DETECT indicated significant multidimensionality. Thus we do not recommend the stand-alone use of DETECT benchmark values to either accept or reject a unidimensional measurement model. However, when DETECT was used in combination with additional indexes of general factor strength and group factor structure, parameter bias was highly predictable. Recommendations for judging the severity of potential model misspecifications in practice are provided.

Keywords: dimensionality assessment, structural equation modeling, bifactor model

In psychological and educational assessments that employ structural equation modeling (SEM), unidimensional measurement models are the desire and de facto the norm. Estimating unidimensional measurement models can be problematic because complex, multifaceted traits are unlikely to yield strictly unidimensional item response data. In reference to the prospect of finding unidimensional data in the real world, McDonald (1981) declared, "Such a case will not occur in application of theory" (p. 102). More recently, Zhang (2007) summarized, "Although it is the most common assumption in the analysis of item response data, the unidimensionality of a set of items usually cannot be met and most tests are actually multidimensional to some extent" (p. 69). Thus, applying a unidimensional measurement model to multidimensional data necessarily causes a mismatch between the model and the data.

Despite this predicament, it is not uncommon to find SEM research that employs a unidimensional measurement model, regardless of the inherent multidimensionality of the data. However, when multidimensional data are forced into a unidimensional measurement model, the resulting model misspecification can lead to seriously biased and potentially misleading parameter estimates (Reise, Scheines, Widaman,

Correspondence should be addressed to Steven P. Reise, Department of Psychology, University of California, Los Angeles, Franz Hall, Los Angeles CA 90095. E-mail: reise@psych.ucla.edu

& Haviland, 2013).[1] For this reason, the consequences of the incongruity between psychological data (which tend to be multidimensional) and commonly applied measurement models (which tend to be unidimensional) have been the subject of much psychometric research, especially in the item response theory (IRT) literature.

In IRT, where the application of unidimensional measurement models dominates, a number of studies have explored the robustness of item parameter estimates to violations of unidimensionality (e.g., Drasgow & Parsons, 1983; Folk & Green, 1989; Kirisci, Hsu, & Yu, 2001). This research has generally shown that if a strong general factor exists in the data, then the estimated IRT item parameters are relatively unbiased when fit to a unidimensional measurement model. Accordingly, in applications of unidimensional IRT models, it is common to see reports of "unidimensional enough" indexes, such as the relative first-factor strength as assessed by the ratio of the first to second eigenvalues (Ackerman, 1989; Embretson & Reise, 2000).

More recently, an appreciable body of literature has focused on the development of more refined nonparametric indexes of the degree to which item response data are "essentially" unidimensional. One measure that has drawn a great deal of attention in IRT is the Dimensionality Evaluation to Enumerate Contributing Traits index (DETECT; Kim, 1994; Zhang, 2007; Zhang & Stout, 1999). The DETECT index attempts to measure the degree of multidimensionality that exists in an item response matrix—the core assumption is that the viability of applying a unidimensional IRT measurement model decreases as the amount of multidimensionality in the data increases. This index and its role in this investigation are discussed in further detail shortly.

In SEM, much less emphasis has been placed on the development and application of statistics that directly index the degree of departure from unidimensionality, even though SEM parallels IRT regarding the potential biasing effects that arise from forcing multidimensional item response data into a unidimensional measurement model. Instead, SEM researchers have traditionally evaluated dimensionality by fitting a unidimensional measurement model and comparing the resulting set of fit index values against established benchmarks (e.g., Hu & Bentler, 1999). Thus, when used to test whether a unidimensional measurement model provides an "acceptable" fit to the data, SEM fit indexes are essentially being used just as DETECT or first-factor strength indexes are used in IRT; that is, SEM fit indexes are used in practice as indicators of whether the data are "unidimensional enough" to avoid serious bias in model parameters.

However, Reise et al. (2013) argued that fit index values can be misleading when they are used to judge the size of the departure from unidimensionality and ultimately the bias in parameter estimates. Although SEM fit indexes might perform well in differentiating between unidimensional and multidimensional data, fit index values are not necessarily prognostic of the degree to which parameter estimates (e.g., structural coefficients) are biased when the model is misspecified. Moreover, some fit indexes are confounded by factors that are irrelevant to dimensionality, such as test length (West, Taylor, & Wu, 2012). In short, although goodness of fit is of course preferable, it is not a direct reflection of unidimensionality, and satisfactory fit indexes do not guarantee that the estimated parameters are unbiased.

Given these facts, we argue that it is important to identify statistical tools that evaluate dimensionality directly, and are therefore potentially more predictive of the degree of parameter bias caused by model misspecification. The nonparametric DETECT index was expressly designed to assess the degree of (multi)dimensionality of an item response matrix, and it has enjoyed widespread use among IRT practitioners. Thus, this study borrows from IRT to determine whether DETECT can be used in SEM as an index of the degree of model misspecification that occurs when fitting multifaceted item response data to a unidimensional measurement model.

DETECT

As an alternative to a general factor strength statistic in IRT, the DETECT index has become increasingly popular. This statistic, developed under the theory of essential unidimensionality (Stout, 1990) assumes that item responses are influenced by a single "dominant" latent dimension (Hattie, Krakowski, Rogers, & Swaminathan, 1996; Nandakumar, 1991; Stout et al., 1996; Zhang & Stout, 1999), and that the residual multidimensionality follows approximate simple structure (i.e., cross-loadings are small). The DETECT index is maximal when the items are appropriately partitioned into homogenous clusters that each measure a subdimension, and the DETECT procedure uses a genetic algorithm to find the clustering that maximizes the index.

Assuming that the sum of raw scores on all the items can serve as a proxy for the common dimension measured by an item set, the DETECT index is based on estimating the covariances between pairs of items, conditioned on the common dimension. It has been shown that, under the assumption of approximate simple structure, items measuring the same secondary dimension beyond the composite have positive conditional covariances, whereas items measuring different dimensions beyond the composite have negative conditional covariances (Mroch & Bolt, 2006; Zhang & Stout, 1999). When item response data are strictly unidimensional, there are no secondary dimensions beyond that reflected in the composite, and thus each item pair has an expected conditional covariance of 0. When a

[1]Moreover, the common practice of "hiding" multidimensionality through the formation of parcels has been subject to much criticism of late (Bandalos, 2002; Meade & Kroustalis, 2006; Sterba & MacCallum, 2010).

test is multidimensional, then the conditional covariance is expected to deviate from zero. The DETECT index can be written as

$$D(P) = \frac{1}{n(n-1)/2} \sum_{i<j} (-1)^{c_{ij}} (\hat{CCov}_{ij} - \overline{CCov}), \qquad (1)$$

where P represents a specific partitioning of items into clusters, n is the number of items, and $\hat{CCov}_{ij}$ is the conditional covariance between items i and j,

$$\overline{CCov} = \frac{1}{n(n-1)/2} \sum_{i<j} \hat{CCov}_{ij}, \qquad (2)$$

and $c_{ij} = 0$ (if items i and j are in the same cluster) or 1 (if items i and j are in different clusters; Mroch & Bolt, 2006; Zhang & Stout, 1999). If approximate simple structure exists, then $D(P)$ will be maximized at the dimensionally homogeneous cluster partition P of a scale, where each cluster represents a different underlying dimension. As Zhang and Stout (1999) summarized, the maximum DETECT value "indicates the amount of multidimensionality the test displays, that is, the size of the departure from being perfectly fitting to a unidimensional model" (p. 219). Low values of DETECT are taken as indicators that the data are essentially unidimensional, and higher values indicate that a multidimensional latent variable measurement model (among other possible remedies) might be required (Roussos & Ozbek, 2006; Stout et al., 1996).

Based on simulation studies, Kim (1994) proposed DETECT benchmarks for unidimensionality (DETECT $<$ 0.1), weak dimensionality (0.1 $<$ DETECT $<$ 0.5), moderate dimensionality (0.51 $<$ DETECT $<$ 1.0), and strong dimensionality (DETECT $>$ 1.0). Stout et al. (1996) simplified this yardstick by stating, "Maximum *DETECT* values near .1 or less are indicative of (possibly essential) unidimensionality (see Stout, 1990), and values greater than 1.0 indicate that sizable dimensionality is present" (p. 348). Most recently, Roussos and Ozbek (2006) refined the original benchmarks by establishing several guidelines based on both simulated and real data analyses:

- DETECT $<$ 0.2 = weak multidimensionality/approximate unidimensionality.
- 0.2 $<$ DETECT $<$ 0.4 = weak to moderate multidimensionality.
- 0.4 $<$ DETECT $<$ 1.0 = moderate to large multidimensionality.
- DETECT $>$ 1.0 = strong multidimensionality.

However, all of these guidelines were based on internal criteria such as the expected values of DETECT under the null condition of unidimensionality (e.g., Finch & Habing, 2005; Gierl, Leighton, & Tan, 2006). In this study, the DETECT index is evaluated in terms of its ability to predict the bias in factor loading estimates, as described next.

THIS STUDY

This study differs from prior work on DETECT in several important ways. First, prior work on DETECT has occurred exclusively in the context of IRT modeling. Nevertheless, due to the mathematical equivalence of IRT and SEM measurement models (Takane & de Leeuw, 1987), we argue that the DETECT index should be equally useful in the evaluation of unidimensional SEM measurement models. In other words, if a researcher wants to apply a unidimensional SEM measurement model to a set of items, but is concerned about the degree to which the data are consistent with this model, then the DETECT index should, in principle, be useful.

Second, most studies of DETECT have used a simple structure correlated factors model as their working representation of multidimensionality (e.g., Gierl et al., 2006; Roussos & Ozbek, 2006; Zhang, 2007). In this research, our working model of multidimensionality is a bifactor structure (Holzinger & Swineford, 1937)—a multidimensional structure wherein measured variables reflect a single common trait (typically the primary dimension of interest to the researcher) and additional orthogonal group factors that typically arise because of homogeneous content groupings (Reise, 2012; Reise, Moore, & Haviland, 2010).[2]

The advantages of fitting a unidimensional SEM measurement model to data generated from a multidimensional bifactor measurement model are twofold. First, we assume that the general factor in the bifactor model—which represents the one common source of variance among all the test items—best represents the latent variable the researcher intends to measure. This allows us to define parameter bias precisely. Specifically, we define parameter bias as the difference between estimated loadings in a unidimensional solution and true loadings on the general factor in the bifactor model. A further advantage is that using a bifactor model as our working representation of multidimensionality allows us to compare the predictive power of DETECT against two bifactor model-based indexes previously shown to be associated with parameter bias in SEM, namely, ECV and PUC (Reise et al., 2013).

In sum, our primary goal is to evaluate the utility of DETECT, originally developed under an IRT framework, to determine the adequacy of fitting a unidimensional measurement model in an SEM context. To our knowledge, research to date has not attempted to directly link DETECT values

[2]Note that this structure is completely consistent with DETECT's assumption of approximate simple structure and previous research because any bifactor structure with one general and 1 – P group factors can be transformed into a simple structure correlated factors model with 1 – P factors.

with parameter bias in either IRT or SEM. Our secondary goals are to (a) further explore the factors that influence DETECT values, and (b) evaluate DETECT in relation to previously researched parametric bifactor model-based indexes of factor strength and structure.

METHOD

Data Simulation Procedure

The Monte Carlo simulations described here were programmed in the R statistical software package (R Development Core Team, 2011). To create a multidimensional population measurement model, we specified the true factor loadings (λ) for a bifactor model with a given number of binary items and P latent factors (one general and $P - 1$ group factors). To create optimal conditions for estimating tetrachoric correlations, we specified that the factor threshold equaled 0.0 for each item. The resulting bias in parameter estimates should be due solely to multidimensionality and not to estimation bias caused by item differences in the threshold. The simulated population model was perfectly bifactor, meaning that each item loaded on both the general factor and one (and only one) group factor, and all factors were orthogonal. To control the direction[3] and degree of bias in the parameter estimates, all group factors were balanced (i.e., the same number of items and the same loadings).

Once the true population bifactor loading pattern was specified, we then computed the implied correlation matrix among the items. Cholesky decomposition was then used to generate continuous normally distributed data based on the true correlation matrix for $N = 10{,}000$ cases. These data were then dichotomized by coding any matrix element greater than 0.0 as a 1, and 0 otherwise. Thus, all items had a proportion endorsed around .50.[4]

Data

A total of 300 distinct data sets of $N = 10{,}000$ were simulated by specifying 15 factor structures and 20 conditions within each structure. The 15 distinct factor structures shown in Table 1 were created to differ in the number of items, the number of group factors, and the number of items per group factor, and thus they differed in the *percentage of uncontaminated correlations* (PUC)—a bifactor model-based index of factor structure.

TABLE 1
15 Data Structures

Structure	*Number of Items*	*Number of Group Factors*	*Items per Group Factor*	*PUC*
1	9	3	3	.75
2	18	2	9	.53
3	18	3	6	.71
4	18	6	3	.88
5	24	2	12	.52
6	24	3	8	.70
7	24	4	6	.78
8	24	6	4	.87
9	24	8	3	.91
10	36	2	18	.51
11	36	3	12	.69
12	36	4	9	.77
13	36	6	6	.86
14	36	9	4	.91
15	36	12	3	.94

Note. $N = 10{,}000$. PUC = percentage of unconfounded correlations.

When data follow a bifactor structure with positive factor loadings, the correlations among the items within group factors are inflated due to both general factor and group factor variance. On the other hand, the correlations among the items in different group factors only reflect variance from the general dimension and are thus uncontaminated by multidimensionality. For each bifactor structure specified in this research, PUC was computed via a simple equation:

$$PUC = \frac{\frac{n_{items}(n_{items}-1)}{2} - n_{groups} \times \frac{n_{ipg}(n_{ipg}-1)}{2}}{\frac{n_{items}(n_{items}-1)}{2}}, \qquad (3)$$

where n_{items} is the total number of items, n_{groups} is the number of group factors, and n_{ipg} is the number of items per group factor. For any fixed number of items, PUC increases as the number of small group factors increases and decreases when there are a few large group factors. For example, an 18-item test will have a larger PUC if there are six 3-item group factors (PUC = .88) rather than three 6-item group factors (PUC = .71).

Within each factor structure, we also varied the relative strength of the general to group factors. Within each of the 15 structures, we specified a completely crossed design with five levels of true loadings on the general factor (.3, .4, .5, .6, and .7) and four levels of true loadings on the group factors (.3, .4, .5, and .6). This minimum average loading was selected because loadings below .3 can be considered inconsequential (McDonald, 1999), and the maximum group loading of .6 was specified to avoid computation problems that can occur when item communalities become too high.

[3]Because of the balanced design, when multidimensional data are forced into a unidimensional model, estimated loadings in the unidimensional model will always be higher than the true loadings on the general factor in the bifactor model. The degree of bias is influenced by other factors, as described later.

[4]Note that we did not vary the sample size because this is not a study of estimation error, but rather a study of DETECT's ability to predict bias in parameter estimates, and the factors that influence its predictive power.

As a consequence of this design, within each data structure we were able to manipulate relative general factor strength, as defined by the *explained common variance* (ECV; Reise et al., 2010; Ten Berge & Sočan, 2004)—the common variance explained by the general factor, divided by the total common variance. Ten Berge and Sočan (2004) described this ratio as "a natural coefficient of 'closeness to unidimensionality'" (p. 621), but here we prefer to interpret ECV solely as a relative general factor strength index:

$$ECV = \frac{\sum \lambda_{Gen}^2}{\sum \lambda_{Gen}^2 + \sum \lambda_{GR1}^2 + \sum \lambda_{GR2}^2 + \cdots + \sum \lambda_{GRn}^2}. \quad (4)$$

In Equation 4 the numerator is the eigenvalue associated with the general factor and the denominator is the sum of the eigenvalues of the general factor and each group factor.

Procedure

For each of the 300 (15 factor structures × 20 parameterizations) data sets of $N = 10{,}000$ dichotomous item responses, we used maximum likelihood based on a tetrachoric correlation matrix to estimate a unidimensional SEM using EQS software (Bentler, 2006). For each data set, we recorded the ECV and PUC based on the true population structure. The DETECT program was used to estimate the degree of multidimensionality based on the maximum DETECT value. For each run, we set the maximum number of factors in DETECT to equal the number of group factors plus three (so as to allow the algorithm to overestimate the number of clusters)[5]. In 291 of the 300 data sets (97%), DETECT was accurate in identifying the correct latent dimensionality, and in assigning items to group factors.[6]

Finally, for each of the 300 conditions, we recorded the average bias in the factor loading estimates by taking the difference between the true loadings on the general factor and the estimated loadings in the unidimensional model, and then averaging. We then computed the relative bias, which we focus on here, by dividing absolute bias by the true general factor loading. Following Reise et al. (2013), we describe the relative bias as trivial if it is less than 5% (strict) or 10% (liberal).

[5]Again, if a bifactor structure has one general and three group factors, DETECT should recognize this as three groups of homogeneous items, and base the computation of the maximum DETECT value on this correct partitioning.

[6]One problem occurred in data structure 9 (24 items, 8 group factors of 3 items) in the .6 general and .3 group loading condition where DETECT identified six dimensions rather than eight. The remaining eight inaccuracies were from data structure 15 (36 items, 12 group factors of 3 items) in the conditions where the group factor loadings were small (.3 and .4). In those conditions, DETECT slightly underestimated the number of latent dimensions as being between 8 and 11, rather than the correct 12. We comment further on these findings in the discussion.

TABLE 2
Structure 5: 24 Items, 2 Group Factors, and 12 Items Per Group Factor (PUC = .52)

λ_{GEN}	λ_{GRP}	*ECV*	*DETECT*	*Absolute Bias*	*Relative Bias*
.7	.3	.84	.87	.03	.04
.6	.3	.80	.84	.03	.06
.5	.3	.74	.77	.04	.08
.7	.4	.75	1.59	.05	.08
.6	.4	.69	1.47	.06	.10
.7	.5	.66	2.69	.08	.12
.4	.3	.64	.78	.05	.13
.5	.4	.61	1.39	.07	.14
.6	.5	.59	2.35	.09	.15
.7	.6	.58	4.15	.11	.16
.3	.3	.50	.76	.06	.22
.4	.4	.50	1.40	.09	.22
.5	.5	.50	2.30	.11	.22
.6	.6	.50	3.58	.13	.22
.5	.6	.41	3.33	.15	.30
.4	.5	.39	2.13	.13	.32
.3	.4	.36	1.36	.11	.36
.4	.6	.31	3.20	.18	.44
.3	.5	.26	2.10	.16	.53
.3	.6	.20	3.24	.21	.71

Note. $N = 10{,}000$. λ_{GEN} = average general factor loading; λ_{GRP} = average group factor loading; ECV = explained common variance; PUC = percentage of unconfounded correlations.

RESULTS

Within Data Structures

There are too many data structures to display the complete within data structure results, but several key points can be made by examining the results for three structures that vary in PUC. Tables 2 through 4 include the results for each combination of general and group factor loadings for data structures 5 (24 items, 2 group factors of 12 items, PUC = .52), 1 (9 items, 3 group factors of 3 items, PUC = .75), and 15 (36 items, 12 group factors of 3 items, PUC = .94), respectively. In each table, ECV, DETECT, absolute bias, and relative bias are shown and the values are arranged in order of relative bias.

There are three key observations from these tables. First, within each condition, ECV and DETECT appear to be related to absolute bias and relative bias. Second, for any given combination of factor loadings, ECV values do not change across model structure (number of group factors and items per group factor) whereas DETECT values do. When the general factor has an average loading of .7 and the group factors have average loadings of .3, for example, ECV is always .84, but DETECT values are .87, .66, and .20 for the three structures. The implication is that, controlling for ECV, DETECT determines that data structures with higher PUC are less multidimensional. Third, and most important, it is clear that both absolute and relative bias decrease as a function of PUC.

TABLE 3
Structure 1: 9 Items, 3 Group Factors, and 3 Items Per Group Factor (PUC = .75)

λ_{GEN}	λ_{GRP}	*ECV*	*DETECT*	*Absolute Bias*	*Relative Bias*
.7	.3	.84	0.66	0.02	0.02
.6	.3	.80	0.61	0.02	0.03
.5	.3	.74	0.75	0.02	0.04
.7	.4	.75	1.30	0.03	0.04
.6	.4	.69	1.29	0.03	0.05
.7	.5	.66	2.17	0.04	0.06
.4	.3	.64	0.85	0.03	0.07
.5	.4	.61	1.18	0.04	0.08
.6	.5	.59	2.01	0.05	0.08
.7	.6	.58	3.38	0.06	0.09
.3	.3	.50	0.90	0.04	0.12
.4	.4	.50	1.30	0.05	0.12
.5	.5	.50	1.83	0.06	0.12
.6	.6	.50	2.88	0.07	0.12
.5	.6	.41	2.73	0.08	0.17
.4	.5	.39	1.80	0.07	0.18
.3	.4	.36	1.39	0.06	0.20
.4	.6	.31	2.68	0.10	0.25
.3	.5	.26	1.97	0.09	0.30
.3	.6	.20	2.62	0.12	0.41

Note. $N = 10{,}000$. λ_{GEN} = average general factor loading; λ_{GRP} = average group factor loading; ECV = explained common variance; PUC = percentage of unconfounded correlations.

TABLE 4
Structure 15: 36 Items, 12 Group Factors, and 3 Items Per Group Factor (PUC = .94)

λ_{GEN}	λ_{GRP}	*ECV*	*DETECT*	*Absolute Bias*	*Relative Bias*
.5	.3	.74	0.20	0.01	0.01
.6	.3	.80	0.18	0.00	0.01
.6	.4	.69	0.33	0.01	0.01
.7	.3	.84	0.20	0.00	0.01
.7	.4	.75	0.34	0.01	0.01
.7	.5	.66	0.56	0.01	0.01
.4	.3	.64	0.22	0.01	0.02
.5	.4	.61	0.33	0.01	0.02
.6	.5	.59	0.52	0.01	0.02
.7	.6	.58	0.89	0.01	0.02
.3	.3	.50	0.24	0.01	0.03
.4	.4	.50	0.34	0.01	0.03
.5	.5	.50	0.49	0.01	0.03
.6	.6	.50	0.77	0.02	0.03
.4	.5	.39	0.51	0.02	0.04
.5	.6	.41	0.73	0.02	0.04
.3	.4	.36	0.37	0.01	0.05
.4	.6	.31	0.71	0.02	0.06
.3	.5	.26	0.53	0.02	0.08
.3	.6	.20	0.73	0.03	0.11

Note. $N = 10{,}000$. λ_{GEN} = average general factor loading; λ_{GRP} = average group factor loading; ECV = explained common variance; PUC = percentage of unconfounded correlations.

Across Data Structures

We next examine the findings when all 300 data sets are considered. Table 5 presents the bivariate correlations for all research factors and indexes. All correlations were significant at $p < .01$, except for the correlation between DETECT and the general factor loading, and the correlations between test length and both absolute and relative bias. We begin by considering the role of relative factor strength (as measured by ECV) and model structure (as measured by PUC) in predicting bias, as suggested in previous research (Reise et al., 2013).

TABLE 5
Bivariate Correlations

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1. λ_{GEN}	1							
2. λ_{GRP}	0	1						
3. ECV	.75	−.64	1					
4. DETECT	$.10^{ns}$	.73	−.39	1				
5. Absolute bias	−.32	.55	−.60	.82	1			
6. Relative bias	−.54	.43	−.70	.60	.93	1		
7. PUC	0	0	0	−.55	−.68	−.54	1	
8. Number of items	0	0	0	−.18	$-.11^{ns}$	$-.09^{ns}$	.15	1

Note. $N = 10{,}000$. λ_{GEN} = average general factor loading; λ_{GRP} = average group factor loading; ECV = explained common variance; PUC = percentage of unconfounded correlations. *ns* indicates correlations that were not significant at $p < .01$.

ECV and PUC. Table 5 shows that ECV and PUC are related to absolute bias ($r = –.60$ and $r = –.68$, respectively) and to relative bias ($r = –.70$ and $r = –.54$, respectively). Figure 1 displays a scatterplot of relative bias as a function of ECV and includes horizontal lines marking 5% and 10% relative bias. It appears that when ECV is above .70, relative bias is below the 10% benchmark and when ECV is above .80, relative bias is less than 5%. However, the evident fan pattern indicates that as ECV decreases, the range in relative bias increases. The degree of relative bias spread for any particular value of ECV is a function of PUC, as demonstrated next.

Figure 2 depicts the average relative bias across the 20 conditions within each model structure as a function of PUC. There is a near-perfect monotonic relation, such that high values of PUC are associated with a very low average relative bias. The four panels in Figure 3 demonstrate the role of PUC in moderating the relation between ECV and relative bias. Each panel displays a plot of the relationship between relative bias and ECV for four levels of PUC: (a) < .53, (b) .68 > and < .78, (c) .85 > and < .89, and (d) > .90. Inspection of these panels shows that the influence of ECV on relative bias depends critically on PUC. At high levels of PUC (i.e., when there are many small group factors), item parameters display little relative bias regardless of the general factor strength. However, when PUC is low (i.e., when there are a few large group factors), relative strength of the first factor is a critical determinant of bias. To examine the importance of these factors in determining bias in our particular simulation study, we estimated a multiple

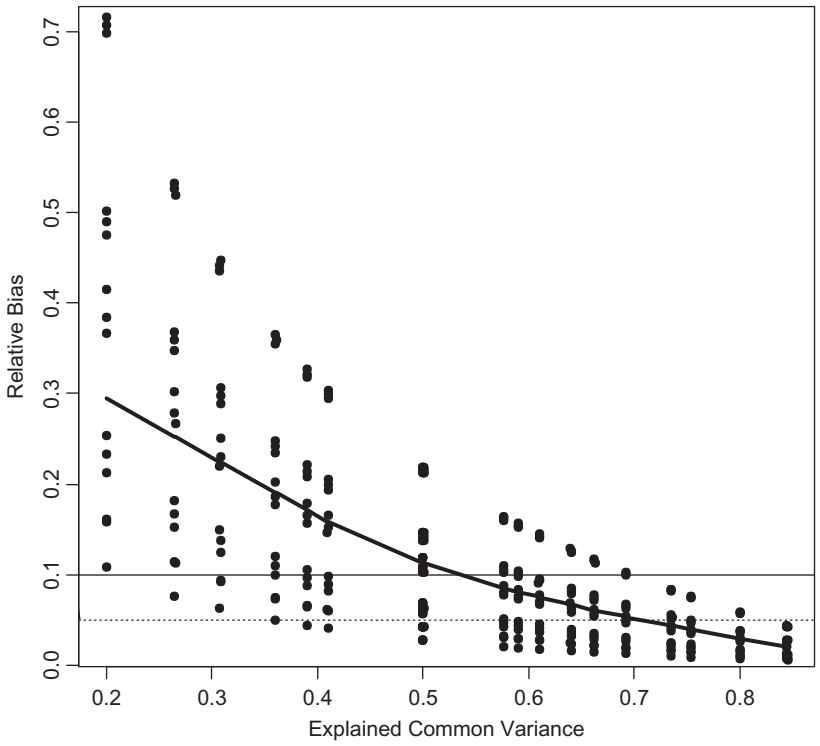

FIGURE 1 Relative bias as a function of explained common variance (ECV).

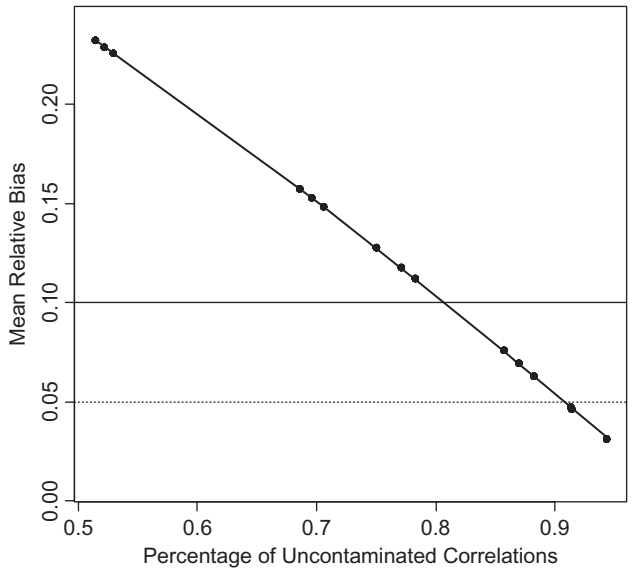

FIGURE 2 Mean relative bias as a function of percentage of uncontaminated correlations (PUC).

regression that included ECV, PUC, and their interaction as predictors of relative bias. The result was an R^2 of .91; this same model predicting absolute bias produced an R^2 of .92. Adding the *DETECT* index did not significantly increase R^2, and the DETECT coefficient was insignificant after including ECV, PUC, and their interaction. Although these results might not generalize to contexts outside of ours, they suggest that relative factor strength, factor structure, and their interaction play a large role in determining bias. With these results serving as a benchmark, we now consider the performance of DETECT.

DETECT. Considered across all data sets, DETECT values were correlated with absolute bias ($r = .82$, $r^2 = .67$), and with relative bias ($r = .60$, $r^2 = .36$). When DETECT is

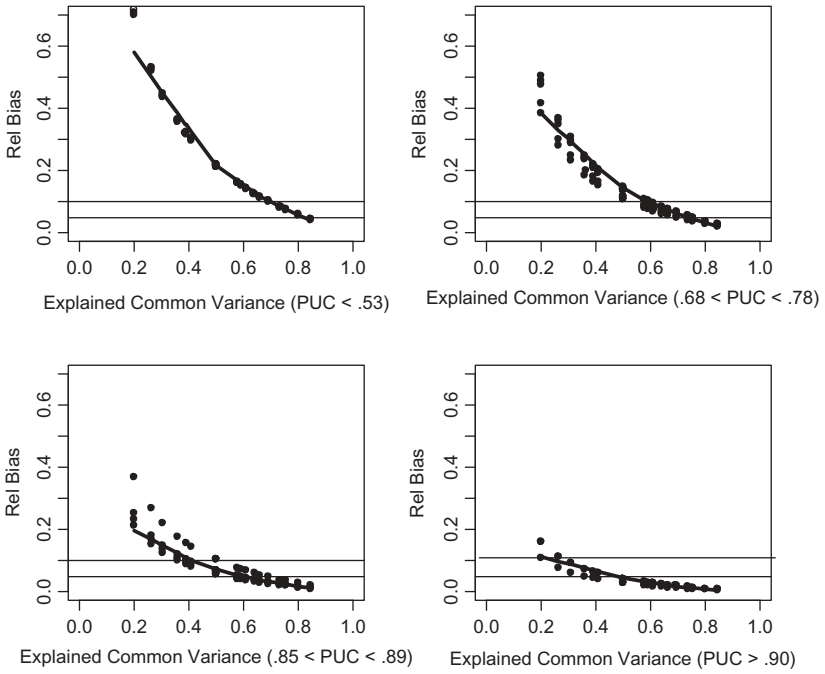

FIGURE 3 The interaction between explained common variance (ECV) and percentage of uncontaminated correlations (PUC) in the prediction of relative bias.

considered on its own, the absolute bias results are superior to the values for either ECV or PUC, but inferior in terms of predicting relative bias. Moreover, DETECT by itself does not perform as well in the prediction of bias as the combination of ECV, PUC, and their interaction, as described previously. To better understand the performance of DETECT under the present conditions, we offer two observations.

First, although DETECT is moderately sensitive to variations in relative factor strength as measured by ECV ($r = -.39$), it is much more sensitive to the size of the average group factor loading ($r = .72$); the size of the general factor loading does not play a significant role ($r = .10$). This is obvious once one recognizes that DETECT is based on the size of the conditional covariances that are found after controlling for the composite score. Because the role of the general factor in the composite score is constant for all conditional covariances, but the role of the group factors varies, the size of the factor loadings on the general factor should not be nearly as important as those of the group factors. In effect, the size of the loadings on the general factor are partialed out in the computation of DETECT. This, in turn, degrades the performance of DETECT in terms of predicting bias. However, a regression combining DETECT, the general factor loading, and the interaction between DETECT and the general factor loading yields $R^2 = .92$ for absolute bias, and $R^2 = .89$ for relative bias—a substantial improvement in predictive power.

Second, as noted previously, ECV is independent of PUC, which itself is an important determinant and moderator of ECV in predicting bias (see Figures 2 and 3). Table 5 shows that DETECT and PUC are correlated ($r = -.55$); as PUC increases, DETECT decreases. The reason these indexes are associated is obvious considering the equations for PUC and DETECT. When PUC increases, there are relatively more covariances between items from different group factors. In turn, the conditional covariances for these item pairs are expected to be negative, thereby contributing to decreased DETECT values. In addition, as PUC increases, there are relatively fewer covariances between items from the same group factor. The conditional covariances for these items are expected to be positive, thereby contributing to higher DETECT values.

The lack of independence between PUC and DETECT has both positive and negative consequences. On the positive side, a virtue of DETECT is that it naturally accounts for the true model structure to some degree and thus, when considered alone, can be more informative than an index such as ECV. On the other hand, the correlation between PUC and DETECT is not perfect, and there are situations where DETECT does not properly account for the true model structure in terms of predicting bias in parameters, as we explore in detail next.

We now consider the specific strengths and weaknesses of using DETECT to judge whether data are too multidimensional for unidimensional modeling. An examination of all 300 data sets revealed that DETECT is always less than 1.0 when the average group factor loading is small (.3), regardless of the strength of the general factor. As the average group factor loading increases in strength, the DETECT values tend to become higher, thus suggesting

multidimensionality. Interestingly, even DETECT values as high as 3.3 can provide parameter estimates with less than 10% relative bias. Alternately, DETECT values as low as 0.6 can show more than 10% relative bias in parameter estimates. These "false negative" and "false positive" scenarios are detailed next.

DETECT false negatives. Figure 4 reveals the data structures that are associated with the "false negatives"—DETECT values greater than 1.0, but relative bias below 5%. For example, the upper right corner of Figure 4 shows that when both general and group factor loadings were strong, data structure 8 (24 items, 6 group factors of 4 items, PUC = .87) had a DETECT value of approximately 1.9, but relative bias around 4.6%. Thus, these values are considered "false negatives"; they would be rejected by the conventional DETECT benchmarks, but the bias in parameter estimates is not substantial. It is important to note that there are no conditions in this sector that have general factor loadings below .6 or group factor loadings below .4. Further, Figure 4 demonstrates that whenever PUC is greater than .70, then the relative bias might be negligible.

These false negatives become even more conspicuous if one accepts a more liberal relative bias level of 10%. The data sets that meet this qualification include conditions in which all group factor loadings as well as general factor loadings are .4 or greater, and PUC values are as low as .51. For example, data structure 1 (9 items, 3 group factors of 3 items, PUC = .75) had an extremely high DETECT value of 3.4, but bias less than 9% when both the general and group factor loadings were high. Similarly, data structures 7 and 12 had DETECT values approaching 3.0, but relative bias of approximately 8%. Clearly, if a researcher allows relative bias as high as 10%, then many data sets with conventionally unacceptable DETECT values will nevertheless provide fairly accurate parameter estimation.

DETECT false positives. Figure 5 shows the "false positive" situations in which DETECT values are generally acceptable (< 1.0), but the relative bias is problematic (> 10%). As indicated by the black shading, the majority of the data points in this plot are characterized by low group factor loadings (.3). In the upper part of Figure 5, for example, one can see that when both general and group factor loadings were very low, data structure 10 (36 items, 2 group factors of 18 items, PUC = .51) had an acceptable DETECT value of approximately 0.75, but high relative bias (∼0.22). It is important to note that all 15 data structures included some combination of general and group factor loadings that resulted in overly biased parameter estimates despite adequately low DETECT values.

DETECT and benchmarks. Finally, Figure 6 focuses on the conditions that satisfy the criteria set forth by Roussos and Ozbek (2006), as mentioned earlier. The first thing to note is that only five data structures are represented in Figure 6 (4, 8, 9, 14, and 15), all of which have high PUC (.87 or higher) and low average group factor loadings (.4 or less). Only three conditions satisfied Roussos and Ozbek's (2006) "approximate unidimensionality" benchmark of DETECT < 0.2, as indicated by the leftmost dotted vertical line in Figure 6. All three conditions were from

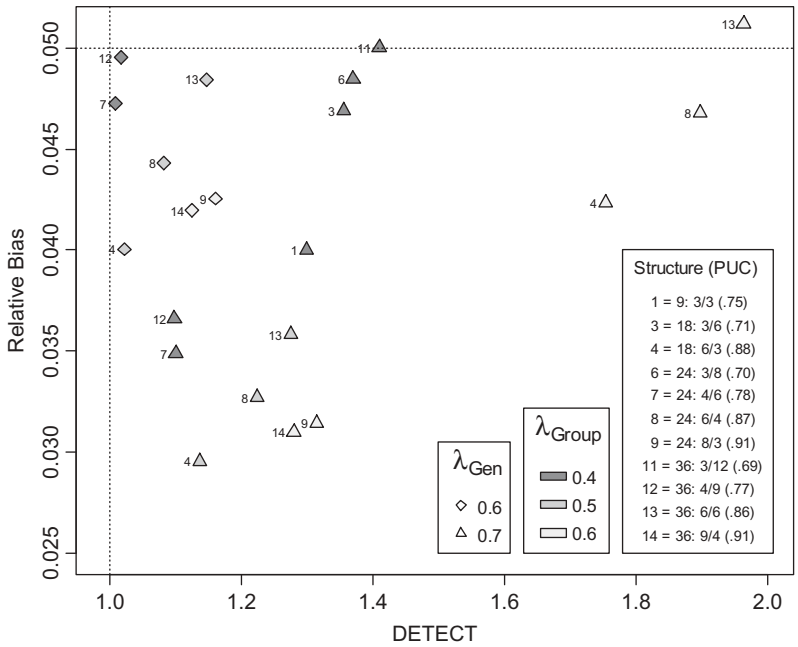

FIGURE 4 False negative DETECT values for conditions with less than 5% relative bias.

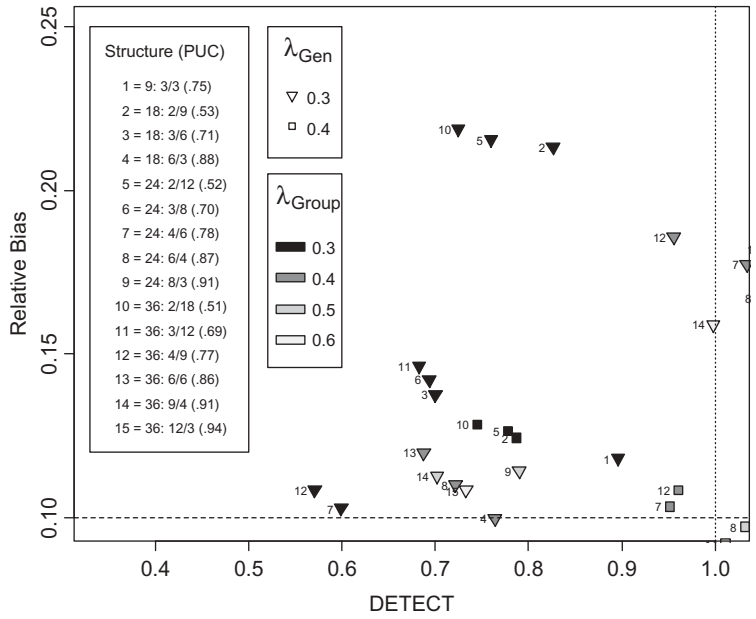

FIGURE 5 False positive DETECT values for conditions with greater than 10% relative bias.

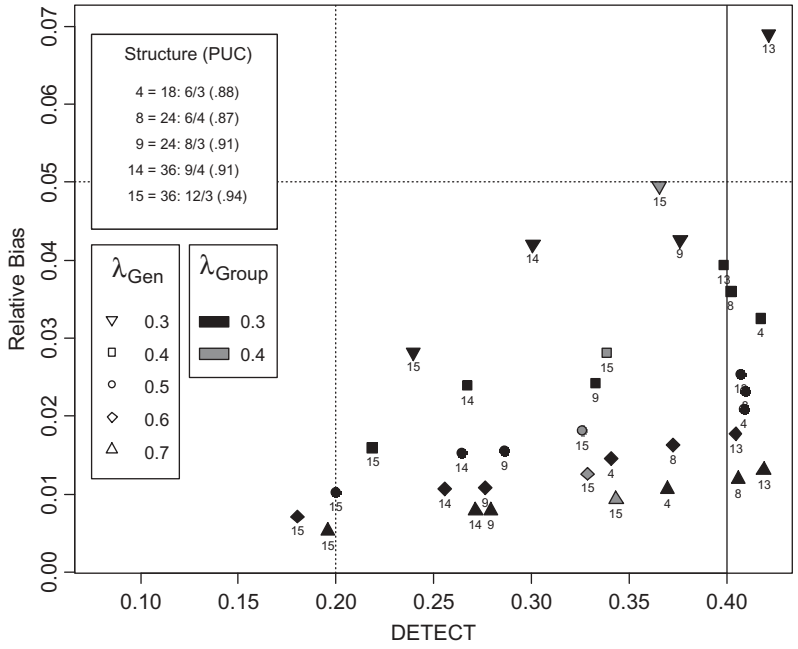

FIGURE 6 Relative bias for conditions that satisfy Roussos and Ozbek's (2006) benchmarks for DETECT.

data structure 15 (36 items, 12 group factors of 3 items, PUC = .94) and had low average group factor loadings of 0.3; they differed only in the strength of the general factor loading (0.7, 0.6, or 0.5). Thus, according to the benchmarks proposed by Roussos and Ozbek (2006), of the 300 data conditions that were simulated, only three were approximately unidimensional.

The area between the two vertical lines in Figure 6 represents Roussos and Ozbek's (2006) DETECT range for "weak to moderate multidimensionality." It is worth noting that structures 14 (36 items, 9 group factors of 4 items, PUC = .91) and 9 (24 items, 8 group factors of 3 items, PUC = .91), as represented by the black upright triangles in the lower center portion of Figure 6, have approximately the same amount of relative bias as the condition that produced the lowest DETECT value. Roussos and Ozbek would identify these data points as weakly multidimensional, despite the fact that they have approximately the same amount of

relative bias (< .01) as the structure 15 conditions that meet their "approximate unidimensionality" criterion.

DISCUSSION

Because item response data are rarely, if ever, strictly unidimensional, researchers have essentially two strategies from which to choose: Ignore the multidimensionality and model the data as unidimensional, or try to model the multidimensionality. In the first case, some degree of bias in parameter estimates is inevitable due to model misspecification. In the second case, specification error in the modeling of multidimensionality will also generally lead to bias. It would be desirable to have some way to evaluate, from theory and data, which strategy is better, or at least to have some way to estimate the amount of bias one is likely to confront with either strategy.

In this research we focused solely on parameter estimation bias caused by misspecifications of the first kind—modeling multidimensional data with a unidimensional measurement model. In particular, we evaluated the utility of the DETECT index for predicting absolute and relative bias in the estimates of factor loadings in a unidimensional SEM model of test items that were, in fact, produced by a multidimensional model. We draw four conclusions.

First, because the DETECT index is based on a specific partitioning of the items (see Equation 1), for the DETECT values to be meaningful, the genetic algorithm used as a first step in this procedure must be precise. In our Monte Carlo simulation, DETECT was able to correctly identify the number of dimensions and to assign items to group factors with 97% accuracy (291 out of 300 data sets). The problematic conditions were mostly associated with the data structure that consisted of a large number of group factors (e.g., 12) and group factor loadings that were small (i.e., .3 and .4). In those conditions, DETECT slightly underestimated the number of latent dimensions. This is an important concern, however, because lengthy achievement tests that are built to measure heterogeneous constructs such as math achievement might very well include numerous small group factors.

Second, previous research (Reise et al., 2013) has argued that relative general factor strength (as measured by ECV) and factor structure (as measured by PUC) are important determinants and moderators of bias in parameter estimates in unidimensional SEM. This finding was replicated in this study by demonstrating that a linear regression predicting relative bias from ECV, PUC, and their interaction produced an R^2 of .91; the same model predicted absolute bias with an R^2 of .92. Moreover, when DETECT was added into this regression model, no significant increase in R^2 was realized. One could argue that these results indicate that once one has addressed general factor strength and structure, DETECT is of little value when estimating the degree of multidimensionality. However, there are clear advantages to DETECT, as we discuss shortly.

Third, in terms of predictive power, *DETECT* was associated with both absolute ($r^2 = .67$) and relative bias ($r^2 = .36$). However, a regression combining DETECT with the general factor loading and their interaction yielded an $R^2 = .92$ for absolute bias and $R^2 = .89$ for relative bias. In short, the flaw of DETECT, in terms of predicting parameter bias, is that its values are based on the size of the conditional covariances (see Equation 1). These conditional covariances are determined by the size of the loadings on the group factor, but they are, by definition, independent of the general factor. However, the size of the general factor, relative to the group factors, is an important determinant of bias. As a consequence, DETECT values must be considered in the context of general factor strength and the structure of the multidimensionality.

Fourth, benchmarks for a single index considered in isolation are not sufficient for predicting bias. Many factors influence parameter bias and different statistics are sensitive to different influences; for that reason, no one statistic will suffice. Even established SEM fit indexes have specific confounds and limitations (West et al., 2012). Similarly, as we have shown, DETECT has a tendency to over- or underpredict relative bias in certain situations. Specifically, DETECT can overpredict relative bias when loadings are strong on both the general factor and the group factors and PUC is at least .70 (see Figure 4). DETECT can underpredict relative bias when both the general and group factor loadings are low, regardless of PUC (see Figure 5).

For multidimensional data of the sort we have studied, we recommend using DETECT in combination with, but prior to, PUC and ECV to predict the relative bias in the estimated factor loadings from a unidimensional measurement model. The DETECT index has the major advantages that (a) although not described here, DETECT yields an index of the degree to which multidimensionality approximates independent cluster structure, and more important, (b) it can be computed directly from a set of data, with no a priori hypothesis about the dimensionality or latent structure involved. This is not true for ECV, or PUC, which requires the estimation of a bifactor model with orthogonal group factor. If DETECT can, with reasonable reliability, provide the number of dimensions and the partition of items into those dimensions, then the results of DETECT can be used to compute PUC and ECV, assuming a bifactor model. DETECT values can then be interpreted in light of these indexes.

Why Not Just Use the Multidimensional Model?

In cases in which the relative bias from fitting a unidimensional measurement model can be expected to be large, one should reasonably consider the second strategy we mentioned earlier—fitting a multidimensional model instead of a unidimensional model. However, two sorts of problems

accompany this strategy. First, even if one is reasonably confident of the structure of the multidimensional model, such a model might include factors that are ill defined or characterized by small two-item local dependencies, thereby increasing the number of parameters that must be estimated and reducing the precision of the estimates.

Much more important, however, is the difficulty in properly specifying the structure of the secondary group factors. Most tests are designed to assess individual differences on a single construct. Yet at the same time, item content heterogeneity can cause small secondary dimensions to emerge (Reise et al., 2010). The impact that these secondary, residual dimensions have on the data is typically substantially smaller than that of the general factor that was intended to be measured; thus, the apparent structure of the secondary dimensions is less likely to be replicable across studies. Even if one fits a model that is misspecified only in the secondary dimensions, then the bias in parameter estimates might well be larger than the bias from fitting a unidimensional model. Again, this is an important question that we hope to research in the future. Many exploratory techniques exist for specifying multidimensionality, but as far as we know, the reliability in specifying a multidimensional model using these techniques is unknown, as is the bias from fitting multidimensional models specified with such exploratory techniques. It certainly seems worth studying how well one can estimate general factor loadings from a multidimensional model specified both correctly, incorrectly, and as a function of some well-posed specification search algorithm.

Limitations

The main limitation of this study relates to the simulated data, which were generated from a pure bifactor model in which the nuisance factors were orthogonal and no cross-loadings were permitted. Further, the use of balanced group factors resulted in bias that was always positive. Of course, such a model is not likely to be found in a real-world situation. However, no highly restrictive multidimensional structure is likely to accurately model item responses drawn from complex psychological measures. Future research should examine bias of the sort we studied under a much broader set of generating models.

CONCLUSION

In SEM, the appropriate use of a unidimensional measurement model requires unidimensional data. Ten Berge and Sočan (2004), however, referred to unidimensionality as a mere hypothesis and stated, "the hypothesis is either beyond verification or it is false" and that "assessing how *close* is a given test to unidimensionality is far more interesting than testing *whether or not* the test is unidimensional" (p. 614). Consistent with this line of thinking, the DETECT "essential unidimensionality" index appears to be useful in an SEM context, and, in our view, is better justified than the uncritical use of fit indexes to judge data dimensionality issues. Nevertheless, when the concern is with parameter bias caused by model misspecification, measuring the degree of multidimensionality does not provide the full picture. For example, in a long test with a reasonably strong general factor and many small group factors, parameter bias is expected to be relatively small regardless of the degree of multidimensionality. Thus, we recommend that DETECT values always be considered interactively with indexes of factor strength (ECV) and factor structure (PUC).

FUNDING

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B080016 to the University of California, Los Angeles. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113–127.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*, 78–102.

Bentler, P. M. (2006). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*, 189–199.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Finch, H., & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, *42*, 149–169.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, *13*, 373–389.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006). Evaluating DETECT classification accuracy and consistency when data display a complex structure. *Journal of Educational Measurement*, *43*, 265–289.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*(1), 1–14.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*, 41–54.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Kim, H. (1994). New techniques for the dimensionality assessment of standardized test data (Doctoral dissertation, University of Illinois at Urbana-Champaign). Retrieved from http://hdl.handle.net/2142/19110

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, *25*, 146–162.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*, 100–117.

McDonald, R. P. (1999). *Test theory: A unified treatment*. London, UK: Psychology Press.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*, 369–403.

Mroch, A. A., & Bolt, D. M. (2006). A simulation comparison of parametric and nonparametric dimensionality detection procedures. *Applied Measurement in Education*, *19*, 67–91.

Nandakumar, R. (1991). Traditional dimensionality vs. essential dimensionality. *Journal of Educational Measurement*, *28*, 99–117.

R Development Core Team. (2011). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667–696.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*, 544–559.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*, 5–26.

Roussos, L. A., & Ozbek, O. Y. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement*, *43*, 215–243.

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research*, *45*, 322–358.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.

Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613–625.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York, NY: Guilford.

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika*, *72*, 69–91.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.