Educational and Psychological Measurement
73(1) 5–26

DOI: 10.1177/0013164412449831
http://epm.sagepub.com

Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective

Steven P. Reise[1], Richard Scheines[2], Keith F. Widaman[3], and Mark G. Haviland[4]

Abstract

In this study, the authors consider several indices to indicate whether multidimensional data are "unidimensional enough" to fit with a unidimensional measurement model, especially when the goal is to avoid excessive bias in structural parameter estimates. They examine two factor strength indices (the explained common variance and omega hierarchical) and several model fit indices (root mean square error of approximation, comparative fit index, and standardized root mean square residual). These statistics are compared in population correlation matrices determined by known bifactor structures that vary on the (a) relative strength of general and group factor loadings, (b) number of group factors, and (c) number of items or indicators. When fit with a unidimensional measurement model, the degree of structural coefficient bias depends strongly and inversely on explained common variance, but its effects are moderated by the percentage of correlations uncontaminated by multidimensionality, a statistic that rises combinatorially with the number of group factors. When the percentage of uncontaminated correlations is high, structural coefficients are relatively unbiased even when general factor strength is low relative to group

[1]University of California, Los Angeles, Los Angeles, CA, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]University of California, Davis, Davis, CA, USA
[4]Loma Linda University, Loma Linda, CA, USA

Corresponding Author:
Steven P. Reise, Department of Psychology, University of California, Los Angeles, Franz Hall, 405 Hilgard Avenue, Los Angeles, CA 90095, USA
Email: reise@psych.ucla.edu

factor strength. On the other hand, popular structural equation modeling fit indices such as comparative fit index or standardized root mean square residual routinely reject unidimensional measurement models even in contexts in which the structural coefficient bias is low. In general, such statistics cannot be used to predict the magnitude of structural coefficient bias.

Keywords

multidimensionality, structural coefficient bias, bifactor model, fit indices

Scientists use latent variable modeling procedures such as structural equation modeling (SEM; Bollen, 1989) or item response theory (IRT; Embretson & Reise, 2000) to operationalize psychological constructs, to scale individual differences, and, when the scientific goal is to explore the relations among psychological constructs with SEM, to control for measurement error. To use classic SEM and IRT estimators, however, researchers often make local independence assumptions that may not hold in their data. In commonly used IRT models, for example, researchers must assume (among other things) that only a single common factor underlies the measured items and produces covariance among them. Similarly, in SEM, researchers often apply a ''unidimensional'' measurement model in which the correlations among indicators are zero after controlling for the latent variable.

In IRT, violations of unidimensionality are referred to as local dependencies among items, whereas in SEM they are referred to as correlated uniqueness. Despite the different descriptors, the consequences of violating local independence are similar: When data have been sampled from a population that does not match the model's assumptions, the latent variable may be improperly identified, such that the item parameters (slopes in IRT, loadings in SEM) cannot be estimated accurately; consequently, modeling applications such as scaling individual differences in IRT or exploring the relations among latent variables in SEM can be inaccurate and misleading.

Researchers know that, in practice, the great majority of measures of complex psychological constructs simply are not locally independent based on a single factor—thus, strictly speaking, measures are not unidimensional. This creates a dilemma; although violations of local independence may result in severe bias levels, they may not. Thus, in IRT, many experts have provided suggestions for deciding whether data are ''unidimensional enough,'' such that model parameters are estimated with limited bias. In fact, the IRT literature is awash with studies of the robustness of IRT model parameter estimates to violations of unidimensionality (Drasgow & Parsons, 1983; Reise, Cook, & Moore, in press; Reise, Morizot, & Hays, 2007), methods for statistically detecting multidimensionality violations (Hattie, 1985; McDonald & Mok, 1995), controlling for multidimensionality (Ip, 2010; Wainer, Bradlow, & Wang, 2007), and judging whether item response data are ''essentially'' unidimensional

(Stout, 1987). These latter ''unidimensional enough'' indices almost always are some function of first-factor strength, such as the ratio of first to second eigenvalues (Embretson & Reise, 2000) or some other index of degree of multidimensionality such as the DETECT procedure (Zhang & Stout, 1999).

In SEM, although the potential biasing effects of forcing multidimensional data into a unidimensional measurement model are well known (e.g., biased parameter estimates can lead to biased structural coefficient estimates), relatively less attention has been given to statistics that directly assess the degree of structural parameter bias resulting from this type of model misspecification. One reason is that, until recently, SEM software allowing for efficient analyses of dichotomous or polytomous item-level data generally was not available. As such, latent variables commonly were represented through item parcels, not items. It seemed odd, however, to perform a thorough ''dimensionality analysis'' for a set of three or four parcels used to define a latent variable. The troubling practice of burying multidimensionality in the item responses in parcels, however, finally is attracting attention (e.g., Bandalos, 2002; Meade & Kroustalis, 2006; Sterba & MacCallum, 2012).

The second reason is that instead of thoroughly examining item-level dimensionality, SEM researchers tend to concentrate on establishing the ''fit'' of the measurement model prior to estimating the full structural model, which includes relations among constructs. Although seldom stated, generally, it is assumed that if the item response data fit the measurement model according to commonly employed goodness-of-fit indices—for example, the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR)—then parameter estimates in the structural model are unbiased, and it is safe to proceed with further model enhancement and evaluation.[1] When the values of these indices are used to judge whether a unidimensional measurement model provides an ''adequate'' fit to the data, essentially they are being used in the same way as ''first-factor strength'' indices in IRT; that is, fit indices are used in practice as indicators that the data are ''unidimensional enough'' to avoid serious bias in model parameters.

In the present article, our primary objective is to examine statistics whose purpose is not to assess whether a measurement model specified as locally independent or unidimensional is misspecified but rather to assess the size of the biasing effect of multidimensionality on structural parameter estimates in an SEM specified as unidimensional. We expect our findings to inform scale construction practice by identifying conditions where the effect of multidimensionality on identifying a common factor is minimal. The secondary goal is to demonstrate that the standard practice in SEM described earlier, one that relies on model fit indices, is insufficient and to suggest a better alternative based on computing factor strength indices. Specifically, drawing from the IRT literature, which emphasizes factor ''strength'' as opposed to model ''fit'' investigations, we conduct a series of demonstrations to clarify the effects of forcing multidimensional data into a unidimensional measurement model in an SEM context.[2]

Our study of model misspecification and parameter bias differs from others in at least two ways. First, our model for multidimensionality is neither the commonly used ''single small secondary nuisance factor'' (e.g., a reading ability dimension influencing responses to several items on a mathematics test) nor the standard ''correlated-traits'' representation (Kirisci, Hsu, & Yu, 2001), so commonly seen in IRT robustness studies. We use a bifactor model—a latent structure where all items load on a general dimension and on one of several group factors (Holzinger & Swineford, 1937). Recently, several authors have argued that a bifactor model is a more realistic representation of complex psychological constructs (e.g., Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Reise, in press; Thomas, 2012). Second, our chief concern is bias in structural coefficients (sometimes called validity coefficients) rather than bias in factor loadings. Of course, we recognize that bias in the former depends in some fashion on bias in the latter. Nevertheless, from an applied perspective, we believe that researchers are more concerned with bias in structural parameters (which play a critical role in their theories) than they are with the specific bias in factor loadings (which rarely are of theoretical interest outside of psychometric reports).

Method

All our populations are determined by a ''perfect'' bifactor model. We did not consider situations in which the group factors were imbalanced, the items loaded on more than one group factor, or the group factors were correlated. In the discussion, we consider the generalizability of our results and address potential limits.

Simulation Procedure

To determine a population measurement model MM_{true} for our general factor *Gen*, we specify the factor loadings (λ) for a bifactor model[3] (which include the relative strength of the factor loadings on *Gen* vs. the factor loadings on the group factors), the number of items, and the number of group factors. The population model was ''strictly'' bifactor (Figure 1B; Holzinger & Swineford, 1937)—with each item loading on the general factor and one and only one group factor and with all factors orthogonal.[4] Importantly, within each of our conditions, the bifactor model had balanced group factors, that is, the same number of items and same loadings.

Next, we specified a criterion latent variable that was measured by three continuous normally distributed indicators. For all conditions, the loadings of the indicators on the criterion latent variable were .60, .65, and .70 (see Figure 1C). These values were selected to represent ''good'' criterion measurement. After specifying a structural path coefficient for the effect of the target latent variable *Gen* on the criterion,[5] we used the complete model to derive an implied population correlation matrix. In the following, we illustrate the results when the structural coefficient was fixed to .50. Our preliminary investigations (not shown) revealed that although the absolute bias in structural coefficient depends on the size of the true coefficient, the relative

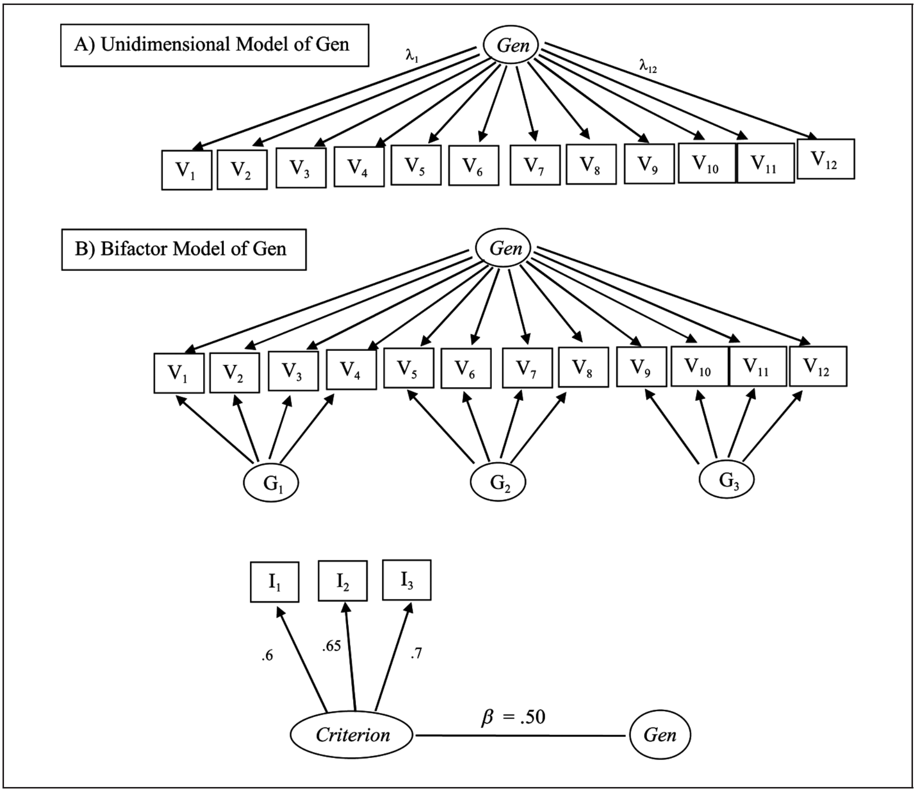

Figure 1. Unidimensional and bifactor measurement models for *Gen* (general factor) and structural model where measured variables are boxed, latent variables are standard normal variates with mean 0 and variance 1, shown in circles

bias (expected estimated value/population value) does not. As a consequence, all the critical points we wish to make here can be achieved using a single criterion value.

Given the true population correlation matrix, the final step involves specifying a unidimensional measurement model for *Gen* (MM_{est}) and estimating the parameters (Figure 1A). EQS (Bentler, 2006) was used for all analyses. For each condition, two analytic models were run. In the first analytic model, only a unidimensional measurement model for the predictor items was specified, so group factors were not estimated. This model was used to record fit indices and factor loadings and reflects a preliminary investigation of whether the model has sufficient fit to proceed to fit a structural model. Because group factors were present in the data-generating model but were not specified in this first analytic model, technically, the model was misspecified. In such cases, we expected that fit indices would not meet typical thresholds for acceptance of the model, implying that the model should be rejected. Because we were interested in assessing structural bias in the presence of model misspecification, the first analytic model provides evidence of the magnitude of misspecification or the magnitude of model misfit to the data on the predictor variables. In the second analytic model, the criterion items and a structural path were included, and estimated structural coefficients were recorded.

Data Structures

In total, six bifactor data structures were specified that varied in the number of items, number of group factors, and number of items per group factor. Specifically, these were as follows:

Structure	Number of items	Number of group factors	Items per group factor	Percentage of uncontaminated correlations
1	9	3	3	.75
2	18	3	6	.71
3	18	6	3	.88
4	36	3	12	.69
5	36	6	6	.86
6	36	12	3	.94

These data structures were designed to vary in ''percentage of uncontaminated correlations'' (PUC). Consider Structure 1, which has nine items, three group factors, and three items per group factor. There are $(9 \times 8)/2 = 36$ unique correlations among nine items. If the structure is bifactor, however, correlations among the items within group factors are affected by two sources of variance (general and group). In Structure 1, there are 3 unique correlations within each group factor times 3 group factors equaling 9 correlations that arise from both *Gen* and the group factor and, thus, are confounded when estimating factor loadings on *Gen* with a model specified as unidimensional. In turn, all correlations among items from different group factors arise solely from the general factor. In Structure 1, there are $36 - 9 = 27$ such correlations. Thus, the PUC for the first condition was $27/36 = .75$. PUC values for the five remaining conditions described above were .71, .88, .69, .86, and .94, respectively. As will be clear shortly, PUC is an important factor in moderating the effects of factor strength on the biasing effects of forcing bifactor data into a unidimensional model.

Within each of the six structures, we specified a completely crossed design with five levels of loadings on the general factor (.3, .4, .5, .6, and .7) and four levels of loadings on the group factors (.3, .4, .5, and .6).[6] The reasoning behind the factor loading values is that we believed that any item set with average loadings of less than

.30 is not worth considering. Moreover, we limited the average loadings for the group factors to .60 (note: this limit also prevented computational difficulties). Thus, the total number of conditions = 6 structures times 5 levels of the general factor times 4 levels of the group for a total of 120 conditions.

Strength Indices

For each condition, two ‘‘factor strength’’ indices were calculated. First, based on the population loading matrix, we computed the *explained common variance* (ECV; see Reise, Moore, & Haviland, 2010, in press; Ten Berge & Socan, 2004), which is the common variance explained by the general factor divided by the total common variance (see Equation 1, where *Gen* is the general factor and *GR1* to *GR3* are three group factors).

$$\mathrm{ECV} = \frac{\sum \lambda^2_{Gen}}{\sum \lambda^2_{Gen} + \sum \lambda^2_{GR1} + \sum \lambda^2_{GR2} + \sum \lambda^2_{GR3}}. \quad (1)$$

The ECV index is a natural and easy-to-interpret index of the degree of unidimensionality, or relative strength, of general to group factors. The ECV can be high whenever there is little common variance beyond the general trait, regardless of the size of the item loadings on the general trait. Also, note that with the present design, ECV values are unaffected by the number of items on a test or PUC.

Second, based on the population loading matrix, for each condition, we computed coefficient omega hierarchical (*omegaH*; McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005; Zinbarg, Yovel, Revelle, & McDonald, 2006) as shown in Equation 2, where *R* is a population correlation matrix and *Gen* is the general factor in a bifactor model. When data are bifactor, relative to ECV, is a more direct index of general factor strength (sometimes referred to as first-factor saturation) than it is an indicator of degree of unidimensionality per se (i.e., how much common variance is due to the general trait).

$$omegaH = \frac{\left(\sum \lambda_{Gen}\right)^2}{\sum R}. \quad (2)$$

Moreover, values of *omegaH* can be interpreted as an estimator of how much variance in summed (standardized) scores can be attributed to the single general factor (McDonald, 1999). All else being equal and assuming that all items load on the general factor, *omegaH* values increase as test length increases; for this reason it is possible for *omegaH* to be very high even in the presence of clear multidimensionality. *OmegaH* values also are affected by PUC as explained in more detail shortly.

Model Fit Indices

For each condition, we also computed the following three fit indices: (a) the SRMR (Bentler, 2006; Hu & Bentler, 1999), (b) the RMSEA (Browne & Cudeck, 1993), and

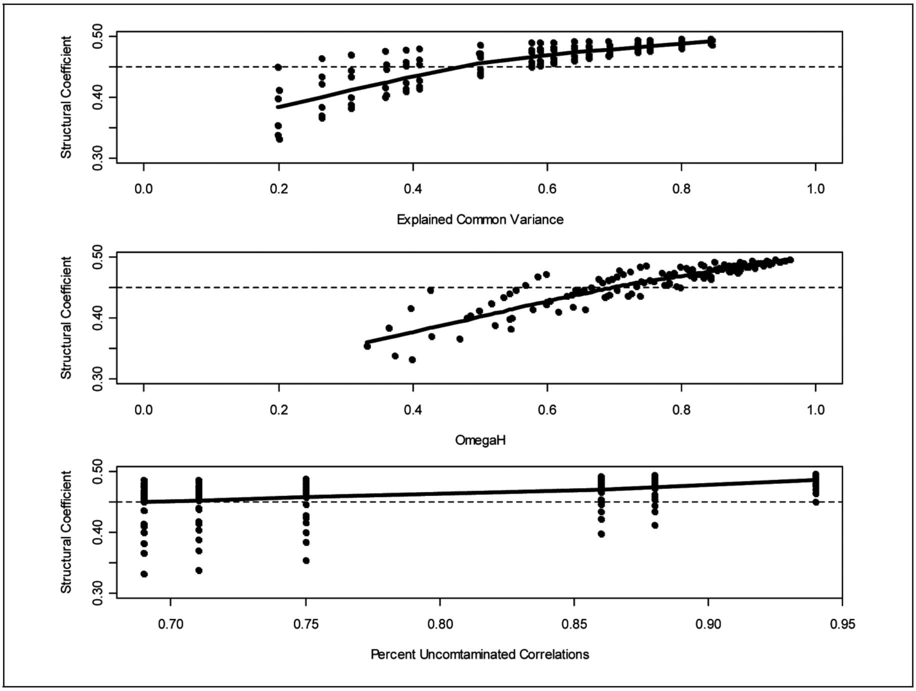

Figure 2. Estimated Structural Coefficient as a Function of ECV, OmegaH, and PUC Across All Conditions.

(c) the CFI (Hu & Bentler, 1999). We chose these three indices for their diversity, their popularity in the research literature (i.e., at least the normal-theory versions), and their endorsement by psychometricians. Various authors have recommended different benchmarks for adequate fit or for distinguishing between correctly specified and misspecified models. Here we use those recommended by Hu and Bentler (1999); namely, RMSEA = .06, SRMR = .08, and CFI = .95. Given that in the population the true factor structures are bifactor and the fitted models are unidimensional, in theory, these indices should indicate that all models are misspecified. Nevertheless, given the known confounds and limitations of these indices (West, Taylor, & Wu, in press), we are not particularly interested in their power to identify that the wrong model has been fitted. Rather, our interest here is whether these fit values are informative as to the degree of structural coefficient bias. In other words, are these popular fit indices good proxies for the degree of structural coefficient bias?

Results

Tables 1 through 6 show the results for each test structure condition. In the bottom row of each table are means within condition. Within each table, results are ordered by ECV values (relative general factor strength) from lowest to highest.

Table 1. Structure 1: 9 Items, 3 Group Factors, and 3 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.33	.19	.41	.14	.45	.35
.3	.5	.26	.37	.13	.53	.10	.42	.38
.4	.6	.31	.48	.21	.50	.14	.53	.40
.3	.4	.36	.40	.08	.70	.06	.39	.42
.4	.5	.39	.52	.14	.65	.10	.50	.42
.5	.6	.41	.60	.24	.57	.14	.62	.43
.3	.3	.50	.43	.04	.89	.03	.37	.45
.4	.4	.50	.55	.09	.80	.06	.48	.45
.5	.5	.50	.64	.16	.72	.10	.59	.45
.6	.6	.50	.70	.29	.59	.14	.71	.45
.7	.6	.58	.78	.38	.57	.14	.80	.46
.6	.5	.59	.74	.19	.75	.10	.69	.46
.5	.4	.61	.68	.10	.85	.06	.57	.46
.4	.3	.64	.59	.04	.94	.03	.46	.47
.7	.5	.66	.81	.24	.75	.10	.78	.47
.6	.4	.69	.77	.12	.88	.06	.67	.47
.5	.3	.74	.71	.05	.95	.03	.56	.48
.7	.4	.75	.84	.15	.88	.06	.77	.48
.6	.3	.80	.80	.06	.96	.03	.66	.49
.7	.3	.85	.86	.08	.96	.03	.75	.49
.5	.45	.54	.63	.15	.74	.08	.59	.45

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .75.

Table 2. Structure 2: 18 Items, 3 Group Factors, and 6 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.37	.15	.41	.16	.44	.34
.3	.5	.26	.43	.11	.53	.11	.40	.37
.4	.6	.31	.52	.17	.49	.16	.52	.39
.3	.4	.36	.49	.07	.68	.07	.37	.41
.4	.5	.39	.58	.11	.62	.11	.48	.41
.5	.6	.41	.64	.19	.54	.16	.60	.42
.4	.4	.50	.64	.07	.77	.07	.46	.44
.5	.5	.50	.69	.13	.68	.11	.57	.44
.6	.6	.50	.73	.22	.57	.16	.68	.44
.3	.3	.50	.54	.03	.89	.04	.34	.44
.7	.6	.58	.79	.29	.54	.16	.77	.45
.6	.5	.59	.77	.15	.71	.11	.66	.46
.5	.4	.61	.74	.08	.82	.07	.54	.46
.4	.3	.64	.69	.03	.93	.04	.43	.46
.7	.5	.66	.83	.19	.71	.11	.75	.47
.6	.4	.69	.82	.10	.84	.07	.64	.47
.5	.3	.74	.79	.04	.94	.04	.53	.48
.7	.4	.75	.87	.12	.84	.07	.73	.48
.6	.3	.80	.86	.05	.95	.04	.62	.48
.7	.3	.85	.90	.07	.94	.04	.72	.49
.5	.45	.54	.68	.12	.72	.10	.56	.44

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .71.

Table 3. Structure 3: 18 Items, 6 Group Factors, and 3 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.50	.13	.36	.11	.36	.41
.3	.5	.27	.54	.08	.52	.08	.35	.43
.4	.6	.31	.65	.14	.47	.11	.45	.44
.3	.4	.36	.57	.05	.73	.05	.33	.45
.4	.5	.39	.68	.09	.65	.08	.44	.46
.5	.6	.41	.75	.16	.55	.11	.54	.46
.3	.3	.50	.60	.01	.98	.03	.32	.47
.4	.4	.50	.71	.05	.83	.05	.42	.47
.5	.5	.50	.78	.11	.72	.08	.53	.47
.6	.6	.50	.83	.19	.58	.11	.63	.47
.7	.6	.58	.88	.25	.55	.11	.73	.48
.6	.5	.59	.85	.13	.75	.08	.62	.48
.5	.4	.61	.81	.06	.87	.05	.52	.48
.4	.3	.64	.74	.02	.98	.03	.41	.48
.7	.5	.66	.90	.16	.74	.08	.72	.49
.6	.4	.69	.87	.08	.89	.05	.62	.49
.5	.3	.74	.83	.02	.98	.03	.51	.49
.7	.4	.75	.91	.10	.88	.05	.71	.49
.6	.3	.80	.89	.03	.97	.03	.61	.49
.7	.3	.84	.93	.05	.97	.03	.71	.49
.5	.45	.54	.76	.10	.75	.07	.53	.47

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .88.

Table 4. Structure 4: 36 Items, 3 Group Factors, and 12 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.40	.12	.42	.16	.45	.33
.3	.5	.26	.47	.08	.53	.11	.41	.37
.4	.6	.31	.55	.13	.48	.16	.52	.38
.3	.4	.36	.55	.05	.68	.07	.37	.40
.4	.5	.39	.62	.09	.61	.11	.49	.41
.5	.6	.41	.66	.14	.53	.16	.60	.41
.6	.6	.50	.74	.16	.55	.16	.69	.44
.4	.4	.50	.69	.06	.76	.07	.46	.44
.5	.5	.50	.72	.10	.66	.11	.57	.44
.3	.3	.50	.63	.02	.91	.04	.34	.44
.7	.6	.58	.80	.21	.52	.16	.78	.45
.6	.5	.59	.79	.11	.69	.11	.66	.45
.5	.4	.61	.78	.06	.80	.07	.55	.46
.4	.3	.64	.76	.03	.93	.04	.43	.46
.7	.5	.66	.84	.14	.69	.11	.75	.46
.6	.4	.69	.84	.08	.82	.07	.64	.47
.5	.3	.74	.84	.03	.94	.04	.53	.47
.7	.4	.75	.89	.09	.82	.07	.73	.48
.6	.3	.80	.89	.04	.94	.04	.62	.48
.7	.3	.84	.92	.05	.93	.04	.72	.49
.5	.45	.54	.72	.09	.71	.10	.57	.44

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .69.

Table 5. Structure 5: 36 Items, 6 Group Factors, and 6 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.54	.11	.33	.12	.38	.40
.3	.5	.26	.60	.07	.47	.09	.35	.42
.4	.6	.31	.69	.12	.43	.12	.46	.43
.3	.4	.36	.65	.04	.69	.05	.34	.45
.4	.5	.39	.73	.08	.58	.09	.44	.45
.5	.6	.41	.78	.13	.49	.12	.55	.46
.5	.5	.50	.82	.09	.65	.09	.53	.47
.3	.3	.50	.70	.01	.99	.03	.32	.47
.6	.6	.50	.84	.16	.52	.12	.64	.47
.4	.4	.50	.78	.05	.78	.05	.43	.47
.7	.6	.58	.88	.20	.49	.12	.74	.48
.6	.5	.59	.87	.11	.69	.09	.63	.48
.5	.4	.61	.85	.06	.82	.05	.52	.48
.4	.3	.64	.82	.01	.97	.03	.42	.48
.7	.5	.66	.91	.14	.68	.09	.73	.48
.6	.4	.69	.90	.07	.84	.05	.62	.48
.5	.3	.73	.88	.02	.97	.03	.51	.49
.7	.4	.75	.93	.09	.83	.05	.72	.49
.6	.3	.80	.92	.03	.96	.03	.61	.49
.7	.3	.84	.95	.04	.95	.03	.71	.49
.5	.45	.54	.80	.08	.71	.07	.53	.47

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .86.

Table 6. Structure 6: 36 Items, 12 Group Factors, and 3 Items per Group Factor

General	Group	ECV	*OmegaH*	RMSEA	CFI	SRMR	Loading	Validity
.3	.6	.20	.67	.08	.37	.08	.33	.45
.3	.5	.26	.70	.05	.57	.06	.32	.46
.4	.6	.31	.79	.09	.49	.08	.42	.47
.3	.4	.36	.72	.03	.85	.04	.31	.48
.4	.5	.39	.81	.06	.69	.06	.42	.48
.5	.6	.41	.86	.11	.56	.08	.52	.48
.3	.3	.50	.75	.00	1.00	.02	.31	.49
.4	.4	.50	.83	.03	.89	.04	.41	.49
.5	.5	.50	.88	.07	.75	.06	.51	.49
.6	.6	.50	.91	.13	.59	.08	.62	.49
.7	.6	.58	.93	.17	.56	.08	.71	.49
.6	.5	.59	.92	.08	.77	.06	.61	.49
.5	.4	.61	.89	.04	.91	.04	.51	.49
.4	.3	.64	.85	.00	1.00	.02	.41	.49
.7	.5	.66	.95	.11	.75	.06	.71	.49
.6	.4	.69	.93	.05	.91	.04	.61	.49
.5	.3	.74	.91	.00	1.00	.02	.51	.49
.7	.4	.75	.96	.06	.90	.04	.71	.50
.6	.3	.80	.94	.01	1.00	.02	.60	.50
.7	.3	.84	.96	.03	.98	.02	.70	.50
.5	.45	.54	.86	.06	.78	.05	.51	.49

Note: ECV = explained common variance; *OmegaH* = omega hierarchical; RMSEA = root mean square error of approximation; CFI = comparative fit index; SRMR = standardized root mean square residual. Percentage of uncontaminated correlations = .94.

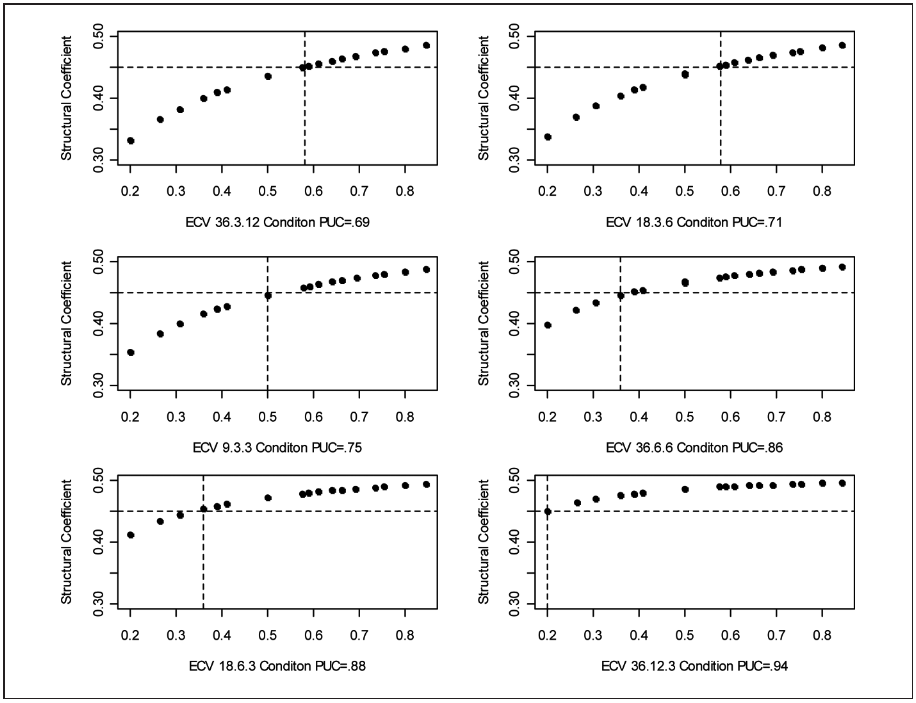

Figure 3. The Interaction Between ECV and PUC in Predicating Structural Coefficient Bias.

We consider first the strength indices within condition and when aggregated across conditions. Figure 2 displays the relation between each strength index (*omegaH*, ECV, and PUC) and the estimated structural coefficient when results are aggregated across the six test structure conditions. In Figure 2, a horizontal dashed line is drawn at structural coefficient equals .45, a 10% bias in the estimate of the structural coefficient, which we take to be the threshold for "severe" bias. This value was taken from Muthén, Kaplan, and Hollis (1987), who note that "bias of less than 10% to 15% could be considered negligible" (cited in Bandalos, 2002, p. 94).

For each structure, there is a monotonic, but nonlinear, relation between ECV and structural coefficient bias—the higher the ECV, the less the bias. This implies that holding the model structure constant, variation in ECV accounts for variation in the bias of the structural coefficient. Across Tables 1 through 6, however, within each general and group factor loading level, ECV values do not change as a function of either test length or test structure; the average level of structural bias, however, does change. Thus, the plot of ECV versus structural coefficient estimates shown in the top panel of Figure 2 is challenging to interpret. On one hand, high values of ECV (>.60) almost always are associated with relatively low bias in the structural coefficient. On the other hand, lower ECV values are related inconsistently to bias in the structural coefficient. Clearly, the relationship between ECV and bias depends on another factor, specifically, the PUC.

To understand the moderating effects of PUC, consider first the bottom panel of Figure 2 showing that the average level of structural coefficient bias decreases as PUC increases. In Figure 3, a set of panels shows the relationship between ECV and structural coefficient estimates within each test structure condition (PUC level). The panels are ordered by PUC value, and within each panel, a horizontal line is drawn at the 10% bias point, and a vertical line is drawn at the lowest ECV value that results in less than 10% structural coefficient bias. Clearly, the effect of ECV on structural coefficient bias is moderated by PUC. For example, in the 9.3.3 condition, ECV values of .50 and higher are associated with validity coefficients of .45 and higher. In the 36.12.3 condition, ECV values as low as .20 are associated with structural coefficients of .45 and higher. Finally, a regression predicting the structural coefficient based on ECV, PUC, and their interaction yielded the parameters shown in Equation 3, with R^2 = .94, $F(3, 116) = 638$, $p < .001$, with all coefficients significant at $p < .01$.

$$\mathrm{B} = -0.0485 + .647(\mathrm{ECV}) + .529(\mathrm{PUC}) - .612(\mathrm{ECV} \times \mathrm{PUC}). \tag{3}$$

We now turn our attention to an alternative strength index, coefficient *omegaH*. Within-condition correlations between ECV and *omegaH* are around $r = .90$, and the correlation is $r = .78$ when data are aggregated across conditions. Unsurprisingly, within each test structure condition, *omegaH* values also are monotonically related to bias in the structural coefficient but not as consistently as ECV. This inconsistency occurs because *omegaH* is more sensitive to the size of the general factor loading, rather than the relative strength of the general to group factors. For example, consider that within each test structure condition, when the general factor loading equals the group factor loading, ECV always is .50, but *omegaH* always is higher for the condition with the larger general factor loading.

Unlike ECV, the average value of *omegaH* changes as a function of test length and particular bifactor structure (PUC). Generally, as test length increases, so does the average *omegaH*, but this is not always the case; consider that the 36.3.12 condition has an average *omegaH* = .72, whereas the shorter 18.6.3 condition has an average *omegaH* = .76. Test structure also plays a role in that, holding test length constant, *omegaH* is relatively higher when more group factors and fewer items per group factor (i.e., the factors that determine PUC). This effect is obvious considering Equation 2. Specifically, the denominator of Equation 2 simply is the modeled raw score variance, and thus Equation 2 could be written as Equation 4:

$$\frac{\left(\sum \lambda_{Gen}\right)^2}{\left(\sum \lambda_{Gen}\right)^2 + \left(\sum \lambda_{GRP1}\right)^2 + \left(\sum \lambda_{GRP2}\right)^2 + \left(\sum \lambda_{GRP3}\right)^2 + \sum_{i=1} \left(1 - h_i^2\right)}. \tag{4}$$

When written this way, it is clear that, all else being equal, fewer and larger group factors result in lower PUC values and relatively lower *omegaH*, and many smaller group factors result in higher PUC values and relatively higher *omegaH*.

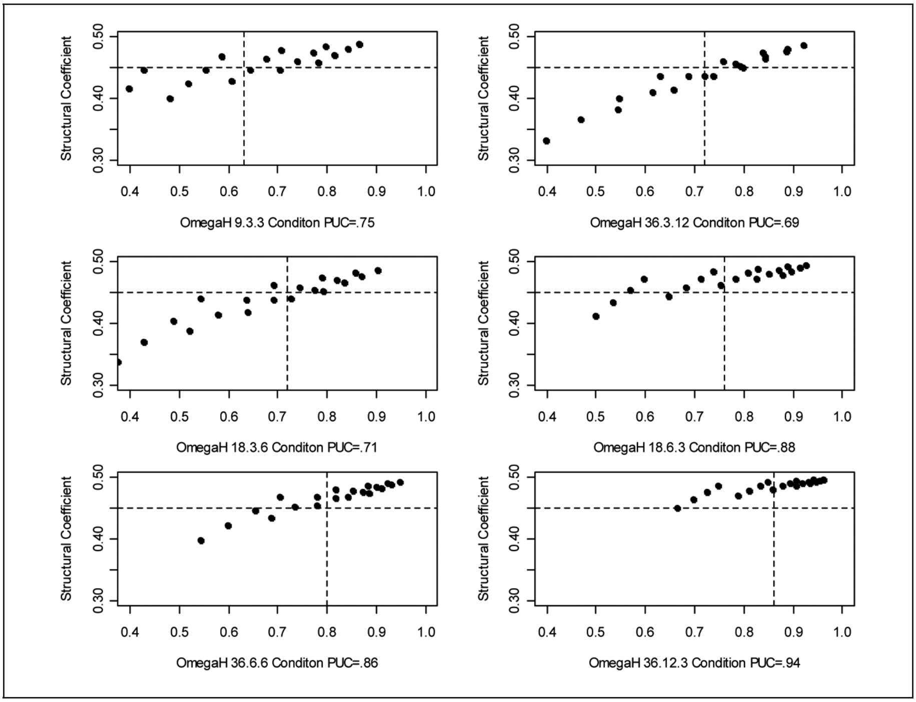

Figure 4. The Interaction Between OmegaH and PUC in Predicating Structural Coefficient Bias.

Despite the fact that the average *omegaH* changes across conditions, the findings in the middle panel of Figure 2 display much the same effect as for ECV; namely, at high values of *omegaH* ($>$.80), structural coefficients display little bias, but at lower values the relation between *omegaH* and the structural coefficient is quite variable. Such findings greatly complicate any attempts to suggest an empirically informed ''benchmark'' for *omegaH* values, and, thus, as earlier, we must consider the interaction between PUC and *omegaH*. Figure 4 displays a set of panels showing the relation between *omegaH* and structural coefficients within each PUC level. However, now the panels are ordered by the average *omegaH* value rather than PUC, and the vertical lines now indicate the average *omegaH* value within condition.

Clearly, both the number of items and PUC play a role in affecting the average *omegaH* value within condition and affect the relation between *omegaH* and structural coefficient bias. When both are high, *omegaH* values tend to become larger with a truncated lower tail (i.e., the minimum value increases). Finally, to discern the unique role of *omegaH* in predicting structural bias, we added *omegaH* into the regression model shown in Equation 3. This resulted in an R^2 value of .95, $F(4, 115) = 518$, $df = 4$ and 115, $p < .001$, and the regression weight for *omegaH* was only $b = .03$ ($p < .01$). The change in R^2 value from model 3 was not

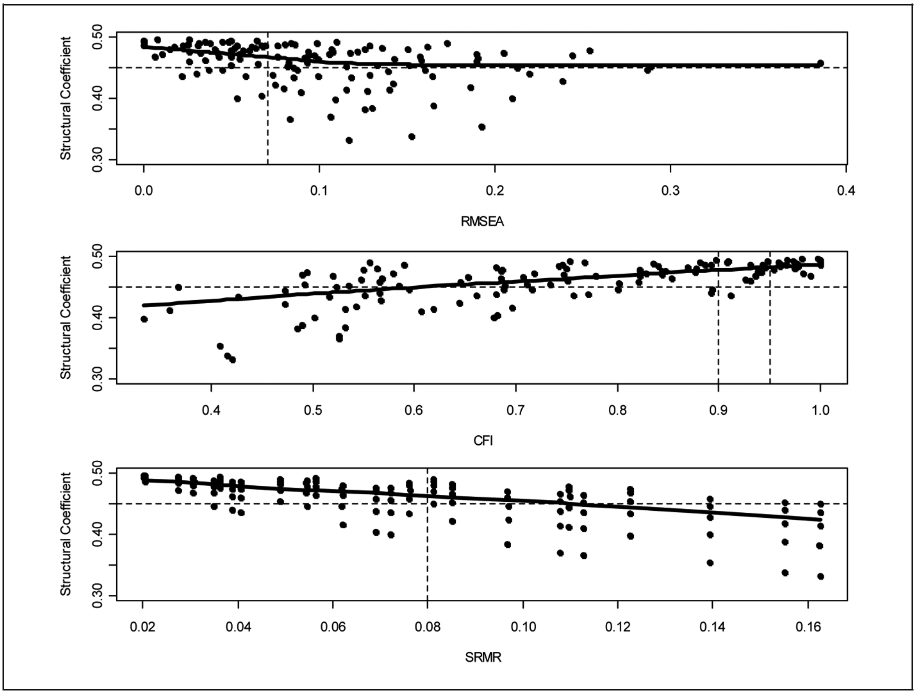

Figure 5. The Relation Between RMSE, CFI, and SRMR With the Estimated Structural Coefficient Across Conditions.

significant, and thus, we conclude that once ECV, PUC, and their interaction are taken into account, *omegaH* makes no further direct contribution to the prediction of bias under the present conditions.

Two findings are of particular note, thus far. First, it appears that relative general factor strength, as assessed by ECV, is more important in influencing structural parameter bias than is general factor saturation, as assessed though *omegaH*. Second, regardless of how it is assessed, ''factor strength'' is of critical importance in determining bias in structural parameter estimates under the present conditions; it by no means, however, tells the entire story. In understanding and predicting structural bias, researchers should consider both factor strength and the data structure and, in particular, the PUC value.

We now consider the performance of practical fit indices. Figure 5 displays three panels showing relations between fit indices and structural coefficient estimates across conditions. As noted, one possible use of such indices is as a guide for indicating when a model is misspecified. In the present case, the unidimensional model is a misspecified model, but the degree of that misspecification varies, arguably as a function of ECV, which also can be thought of as a degree of unidimensionality index. Nevertheless, for each fit index, many models that are wrong in the population clearly are deemed ''acceptable'' using standard benchmarks. In terms of the present study, this is not a concern because we have no interest in judging these indices by their ability to detect that multidimensional (bifactor) data have been fit to a unidimensional model.

The salient question here is whether the values of fit indices can be used to project bias. The linear correlations between fit and structural coefficient are −.30, .67, and −.67 for RMSEA, CFI, and SRMR, respectively. These are in contrast to the r = .77 for ECV and r = .88 for *omegaH* considered across conditions. In the top panel of Table 5, unacceptable RMSEA values ranging from .10 to .20 are sometimes coincident with structural coefficient estimates with relatively small bias, but, under some conditions, acceptable values of RMSEA are coincident with severe bias in structural coefficient. This is not too surprising in the present conditions given that RMSEA is only weakly associated with factors that determine structural coefficient bias, such as the strength indices ECV (r = −.24) and *omegaH* (r = −.12), and mean RMSEA changes little with changes in PUC values. A regression equation including RMSEA, PUC, and their interaction resulted in an R^2 value of only .29—arguing against any attempt to interpret its value as a ''unidimensional enough'' index.

On the other hand, CFI and SRMR appear more promising as ''unidimensional enough'' indices. This is partially because each of these indices is related to the factors that influence bias. Across conditions, CFI is correlated r = .80 with ECV, r = .49 with *omegaH*, but essentially unrelated to PUC (r = .09). SRMR is correlated r = −.57 with ECV and negatively correlated to approximately the same extent with *omegaH* (r = −.41) and PUC (r = −.40). In fact, consider that in Table 6 where PUC is .94, SRMR is no higher than .08 in any condition. Still, a regression predicting the structural coefficient combining CFI, PUC, and their interaction resulted in an R^2 of .74. A similar regression combining SRMR, PUC, and their interaction resulted in an R^2 of .52. These values are far below the R^2 of .94 found when ECV, PUC, and their interaction were used to predict the value of the structural coefficient. Hence, although both CFI and SRMR are somewhat related to factor strength and structural coefficient bias under the present conditions, their values are not as prognostic as are the strength indices, especially ECV and PUC.

Discussion

It is easy to show that in SEM, assuming multivariate normality and when data perfectly match the measurement model, factor loadings and structural coefficients can be estimated in an unbiased way. Thus, when item response data are perfectly unidimensional (local independence with one common factor) and a unidimensional measurement model is used, bias is not a pressing concern. Likewise, when item response data are multidimensional and strictly bifactor and a bifactor measurement model is applied, the general factor will properly reflect the common variance among all the items and parameter estimates will be unbiased.

The challenge of applying SEM, however, is that both the unidimensional and bifactor measurement models are ideal population structures placing very heavy restrictions on data (Saris, Satorra, & van der Veld, 2009). In the real world of psychological data, we do not expect that either of these ideal representations is true in the population (see MacCallum & Tucker, 1991, for details). Rather, given the standard way that many measures of complex psychological constructs are created (Clark & Watson, 1995), we expect that all items will be influenced by at least one common trait. However, because of clusters of relatively more content homogeneous items, we also expect that additional factors will be needed to fully account for item covariances. That is, some type of multidimensional measurement model may be necessary to account adequately for all local dependencies in the data.

When local independence violations are small, they are readily addressable through specifying correlated residual terms. In cases where multidimensionality clearly is at issue, such as when a set of items reflects the measurement of a single general construct but also several group factors, alternative multidimensional measurement models such as second-order or bifactor structures may be needed (Chen, West, & Sousa, 2006). Specifying these more complicated models can create more problems than they solve, however. Thus, given that item response data are rarely strictly unidimensional, and given our belief that in many circumstances, a researcher would much prefer to specify a simple unidimensional measurement model, statistical approaches are needed for deciding when data are ''unidimensional enough,'' such that if a unidimensional model were applied, structural parameter bias is not too severe. As noted in the introduction, many statistical approaches for detecting local independence violations have been proposed in the IRT literature, a wealth of research has been published exploring the robustness of unidimensional model parameter estimates to various forms of multidimensionality violations, and several commonly used ''unidimensional enough'' indices are recommended to judge the acceptability of the data for unidimensional modeling.

In considering the application of a unidimensional measurement model in SEM—perhaps, partially because of the prevalent use of parcels—typically, item responses are not examined in terms of dimensionality or general factor strength. Rather, a unidimensional model may be fit, and if its fit is deemed ''unacceptable,'' a researcher would reject the model and search for some better fitting multidimensional representation. Indeed, some researchers have taken the logic of SEM evaluation and suggested that SEM fit indices be used to evaluate whether item response data are unidimensional enough for IRT modeling (e.g., Cook & Kallen, 2009; Reeve et al., 2007; see also the original work of McDonald & Mok, 1995).

This is unfortunate. The use of fit indices to make judgments about whether the data are unidimensional enough in either IRT or SEM is not optimal if the data have a multidimensional bifactor structure. As shown here, factor strength indices are much easier to link to parameter bias relative to model fit indices. Consider that in the present study of several population structures in which no sampling error was present, several concerns arose with popular fit benchmarks. First, all unidimensional models applied here were misspecified, but in many cases fit values judged the model to be adequate. Thus, as statistics capable of identifying the wrong model under a variety of conditions, these indices performed poorly. Values of fit indices, especially CFI and SRMR, are somewhat related to factor strength and, thus, to parameter accuracy, but the relation is inconsistent.

In contrast, factor strength indices can be linked directly to structural parameter bias under the conditions studied here. Specifically, negative bias in predictive validity coefficients increases as the loadings on the general factor in the bifactor model decrease and the loadings on the group factors increase. In the IRT literature, research has long shown that parameters are accurate if there is a ''strong common trait'' underlying item responses. Our results generalize this finding to SEM by demonstrating that when data are more complex than the measurement model specified, structural coefficient accuracy also appears to depend on the presence of a ''strong common trait.'' Specifically, to the degree that data lack a strong common trait, factor loadings are overestimated and measurement error is underestimated. Not surprisingly, this produces negative bias in structural coefficients.

A second finding was that both the ECV and PUC can be directly related to bias in structural coefficient estimates. Thus, one of our recommendations for practitioners is that if unidimensionality is in doubt (i.e., exploratory analyses suggest the presence of secondary nuisance dimensions caused by clusters of items with similar content), they should estimate a plausible alternative exploratory bifactor model (e.g., a Schmid–Leiman solution; Schmid & Leiman, 1957) and report the values of PUC, ECV, and omegaH. To the extent that PUC is high ($>.80$), the values of the strength indices are less important in predicting bias. When PUC is lower than .80, researchers may consider ECV values greater than .60 and *omegaH* values greater than .70 as tentative benchmarks. Clearly, however, more work is needed to generalize the findings here to a wider variety of conditions.

These conclusions point to two practical suggestions. First, if a researcher suspects multidimensionality, and that multidimensionality takes a bifactor form, we recommend the computation of PUC and at least ECV in addition to fit indices to better inform the consequences of forcing multidimensional data into a unidimensional measurement model. Second, in scale construction, if the goal of the measure is to assess a single individual difference variable, but multidimensionality in response arises because of the inclusion of construct-relevant clusters of diverse items, PUC should be maximized. This is accomplished by having many small group factors, items that are pure indicators of the general factor, and the avoidance of items that cross-load on the group factors. When PUC is high, loading estimates in the unidimensional model should be close to the true loadings on the general factor in the bifactor model. In turn, structural coefficients should be estimated more accurately.

Relative Bias and Its Consequences

Earlier, we cited research that suggested relative bias between 10% and 15% was negligible. Accordingly, in several of our relative bias plots, we identified the 10% line and noted some conditions where average relative bias was within this limit.

Although we recognize the need to provide editors, reviewers, and the general research audience with benchmark values, we do not believe that it is possible when it comes to parameter bias. Stated simply, the issue is too context dependent, and, as found here, absolute bias depends on the size of the predictor–criterion relation but relative bias does not.

Consider a case where a researcher is interested in the relation between a single predictor and criterion, and the true relation is .50. Now, because of unmodeled multidimensionality in the measurement model, the observed estimate is .45, which corresponds to 10% relative bias. We can sympathize with researchers who would respond with little concern, arguing that the relation clearly is identified despite the bias. But now consider cases where many paths are estimated in a model, some representing indirect or mediated effects, others serving as control variables, and so forth. In these more complicated models, even the smallest degree of bias can distort relations among the system of variables, thus rendering all paths potentially inaccurate. In sum, it is wise to minimize bias in parameter estimates when possible. When not possible, the consequences of bias need to be considered within the context of the study goals.

Study Limitations

In the true model for bifactor data, all items loaded on the general factor and only one group factor (no cross-loadings on group factors allowed), the correlation among general and group factors was zero, and the size and strength of the group factors was balanced so not all possible bifactor structures were included. Nevertheless, these MM_{true} models allowed us to study the effects of model misspecification on important parameter estimates as demonstrated here.

Ultimately, however, it is fair to ask: To what extent do the present findings apply to real-world measures, where the covariance structures are much messier (e.g., group factors unbalanced and possibly correlated, doublets causing correlated errors, and so forth)? First, it is impossible to simulate all possible ''real-world'' modeling violations in a single study. Second, we are now developing a general method to allow researchers to study potential consequences of a wider variety of different model violation types. Nevertheless, we maintain that our overall conclusions, in fact, are generalizable—unmodeled multidimensionality in the measurement model causes parameter bias. The specific form and size of such bias, ultimately, will depend on the specific form and size of the modeling violation.

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 is an understandable practice in that if the data are strictly unidimensional and a unidimensional model is fit, or if the data are perfectly bifactor and a bifactor model is fit, it is easy to show that, indeed, parameter estimates are unbiased. Thus, it follows that if a specific model is judged an adequate fit to a set of data, it also should be safe to proceed with fitting a structural model. The fault in the logic is the underlying assumption that ''adequate fit'' implies parameter accuracy.
2. We address the topic of fitting multidimensional models to multidimensional data in the discussion. For now, we assume simply that unidimensional measurement models are used to represent a single construct. This, by far, is the most common practice.
3. All latent variables are normally distributed with mean = 0 and variance = 1.
4. Each measured variable is assumed to be a linear function of its immediate parents (shown in the figures) and Gaussian noise (not shown).
5. As *Gen* and *Criterion* are made to have standard normal distributions, the path coefficient β reflecting the linear dependence of *Criterion* on *Gen* also is the implied correlation between *Gen* and *Criterion*.
6. Small random values were added or subtracted to prevent linear dependencies.

References

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 6.1 Structural equations program manual*. Encino, CA: Multivariate Software.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*, 219-251.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality-of-life. *Multivariate Behavioral Research, 41*, 189-225.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cook, K. F., & Kallen, M. A. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumptions. *Quality of Life Research, 18*, 447-460.

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.

Gustafsson, J. E., & Aberg-Bengtsson, L. (2010). Unidimensionality and the interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs* (pp. 97-121). Washington, DC: American Psychological Association.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Holzinger, K. J., & Swineford, R. (1937). The bifactor 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.

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

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

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502-511.

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.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Erlbaum.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30*, 23-40.

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*, 431-462.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., . . . Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). *Medical Care, 45*(Suppl. 1), S22-S31.

Reise, S. P. (in press). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*.

Reise, S. P., Cook, K. F., & Moore, T. M. (in press). A direct modeling approach for evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. Reise & D. Revicki (Eds.), *Handbook of applied item response theory in typical performance assessment*. New York, NY: Taylor & Francis.

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., Moore, T. M., & Haviland, M. G. (in press). Applying unidimensional item response theory models to psychological data. In K. Geisinger (Ed.), *APA handbook of testing and assessment in psychology*. Washington, DC: American Psychological Association.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Medical Care, 16*, 19-31.

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*, 561-582.

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

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. (1987). A nonparametric approach to assessing latent trait dimensionality. *Psychometrika, 52*, 589-617.

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

Thomas, M. L. (2012). Rewards of bridging the divide between measurement and clinical theory: Demonstration of a bifactor model for the Brief Symptom Inventory. *Psychological Assessment, 24*, 101-113.

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

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

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*, 231-249.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_H. *Applied Psychological Measurement, 30*, 121-144.