Psychological Test and Assessment Modeling, Volume 63, 2021 (1), 95–117

On the Performance of Multiple-Indicator Correlated Traits-Correlated (Methods – 1) Models

Christian Geiser[1] *& Trenton G. Simmons*[2]

Abstract

We examined the performance of two versions of the multiple-indicator correlated traits-correlated (methods – 1) [CT-C(*M* – 1)] model (Eid et al., 2008) in terms of convergence, improper solutions, parameter bias, standard error bias, and power to detect misspecified models. We also studied whether Yuan et al.'s (2015) correction procedure for the maximum likelihood chi-square model fit test yields accurate Type-I error rates and adequate power for these models. The models performed well except for underestimated standard errors for some parameters in specific small-sample conditions. Yuan et al.'s (2015) chi-square correction worked well for correctly specified models but showed limited power to detect misspecified models in small-sample, low-reliability conditions. We recommend that researchers using these models in smaller samples select highly reliable indicators.

Keywords: method effects, multitrait-multimethod (MTMM) analysis, multiple-indicator CFA-MTMM models, CT-C(*M* – 1) model, model-size effect.

[1]*Correspondence concerning this article should be addressed to:* Christian Geiser, Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, UT 84322-2810. E-mail: christian.geiser@usu.edu
[2]Department of Psychology, Utah State University

Multitrait-multimethod (MTMM) analysis (Campbell & Fiske, 1959) is frequently used to study construct validity in the social sciences. The MTMM approach uses a measurement design in which multiple traits T are assessed with multiple methods M (e.g., different rater types, tests, or physiological measures), resulting in $T \times M$ measured variables (e.g., items, test score variables, questionnaire ratings, physiological scores).[3] The convergent and discriminant validity of different measures can be assessed based on the intercorrelations of the variables included in the MTMM design. Correlations between variables that share the same trait but not the same method (so-called *monotrait-heteromethod* correlations) can be used to examine the degree of convergent validity across methods, with high correlations indicating strong convergent validity (agreement between methods). Correlations between measures that do *not* share the same trait can be used to examine discriminant validity. High correlations between measures of different traits within the same method (so-called *heterotrait-monomethod* correlations) indicate a lack of discriminant validity that is potentially due to the presence of shared method (e.g., halo) effects.

Modern approaches to MTMM analysis typically use models of confirmatory factor analysis (CFA) because of their ability to (a) correct for measurement error (unreliability) in the measures, (b) separate variance components due to trait, method, and error influences, and (c) relate method factors to one another as well as to external variables. In the present study, we examined the performance of two multiple-indicator CFA-MTMM models through simulations.

CFA-MTMM Models

Early CFA-MTMM models were based on a single measured variable Y_{tm} (t = trait, m = method) per trait-method combination as in Campbell and Fiske's (1959) original MTMM design with $T \times M$ variables. Single-indicator CFA-MTMM models have been described in detail by, for example, Eid (2000); Lance, Noble, and Scullen (2002); Marsh (1989); and Widaman (1985).

Marsh and Hocevar (1988; see also Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Marsh, 1993) pointed out limitations of single-indicator CFA-MTMM models. Perhaps the most serious problem of single-indicator models is that most of them include only one

[3] From the perspective of latent state-trait (LST) theory (Steyer, Mayer, Geiser, & Cole, 2015), the term "trait" implies a construct that is solely driven by person-specific influences with no situational variation. In cross-sectional data, trait (person-specific) influences cannot be separated from situation or person-situation interaction influences. Therefore, LST theory would consider the constructs studied in cross-sectional MTMM analysis to be *states* (a blend of person, situation, and interaction influences) rather than pure *traits*. In contrast, in classical MTMM analysis, the term "trait" refers more broadly to any construct that a researcher may study, regardless of whether these constructs are trait-like (person-driven) or state-like (situation-driven). Given that the term "multi*trait*-multimethod" is very well-established in the literature, we decided to use it in the present work. At the same time, we note that the traits considered in cross-sectional MTMM analysis are likely to contain both trait and state components. MTMM modeling extensions for longitudinal data (e.g., Geiser, Hintz, Burns, & Servera, 2019) allow disentangling these components.

general method factor per method, implying homogeneous (unidimensional) method effects across different traits for a given method. As a result, method effects are assumed to correlate perfectly (1.0) across different traits (e.g., biases associated with self-reports of extraversion would have to be perfectly correlated with the self-report biases for conscientiousness and the self-report biases for agreeableness).

In a recent study, we found that the unidimensional-method assumption is not supported by empirical applications that have tested this assumption. In Geiser and Simmons (in press), we conducted a systematic review of 20 applications of different multiple-indicator CFA-MTMM models that reported a total of 111 different-trait, same-method method factor correlation estimates. In that study, we found that over 90% of the correlations were < |.90|. The average correlation was |.52|, with most estimates falling into a range between about |.30| and |.60|. Only one single correlation estimate was equal to 1.0. We were thus unable to find even a single application that would have been in line with the assumption of perfect generalization of method effects across traits for all methods—an implication of single-indicator models. When the assumption of perfectly general method effects is violated, convergent validity tends to be overestimated in single-indicator models, whereas method effects and reliability coefficients tend to be underestimated (Geiser & Simmons, in press).

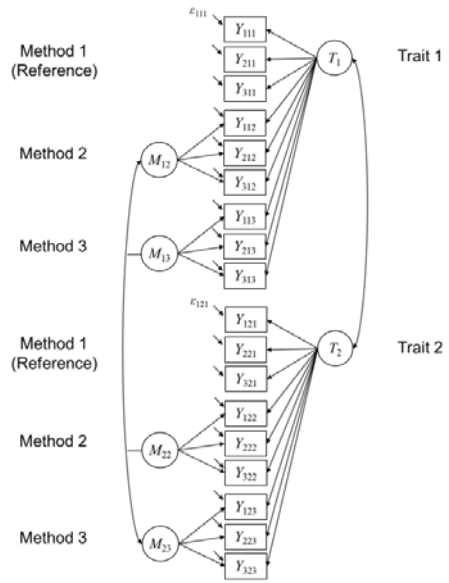

Figure 1:
Multiple-indicator correlated traits-correlated (methods – 1) [CT-C(*M* – 1)] model with general trait factors for a 2 traits × 3 methods design and three indicators per trait-method combination.

Multiple-Indicator Correlated Traits-Correlated (Methods – 1) Models

Marsh and Hocevar (1988; Marsh, 1993) were the first to introduce multiple-indicator extensions of CFA-MTMM models in which each trait-method unit is represented by two or more measured variables Y_{itm} (i = indicator/observed variable/measure; $i = 1, \ldots, I$). This results in a design with $T \times M \times I$ measured variables. As a result of this extended design, there is enough information to identify separate method factors for each trait (so-called *trait-specific* method factors). Eid, Lischetzke, Nussbeck, and Trierweiler (2003) developed the multiple-indicator correlated traits-correlated (methods – 1) model [CT-C($M-1$) model; see Figure 1] that allows researchers to examine trait-specific method factors as well as estimate correlations between these factors to study the degree of generalization of method effects across traits. The multiple-indicator CT-C($M-1$) model thus does not require the assumption of perfectly general method effects within each method.

One advantage of the CT-C($M-1$) approach is that the latent trait and method factors in this approach are explicitly defined based on conditional expectations of measured variables. As a result, all latent variables have a clear psychometric meaning and interpretation based on concepts of classical psychometric test theory (for details, see Eid, 2000; Eid et al., 2003; Geiser, Koch, & Eid, 2014; Koch, Eid, & Lochner, 2018). Other approaches in this tradition were presented by Pohl, Steyer, and Kraus (2008) as well as Pohl and Steyer (2010). In the present article, we focus on the CT-C($M-1$) approach as it was the most-used approach in our recent literature review of applied multiple-indicator CFA-MTMM studies (Geiser & Simmons, in press).

The CT-C($M-1$) model uses a "gold standard" or reference method approach in which $M-1$ (non-reference) methods are contrasted against a reference method (Method 1 in Figure 1). No method factors are included for the reference method so that the trait factors are defined by the reference method. For example, van der Ende, Verhulst, and Tiemeier (2020) studied internalizing and externalizing problem behavior in adolescents using self-reports, parent reports, and teacher reports and used self-reports as reference method. Van der Ende et al. (2020) thus contrasted the self-ratings against parent and teacher ratings. Greiff, Fischer, Wüstenberg, Sonnleitner, Brunner, and Martin (2013) examined the convergent validity of different methods for measuring complex problem solving and used a specific task as reference method. Mazzetti, Schaufeli, and Guglielmi (2018) studied workaholism and work engagement with self- and coworker reports using self-reports as reference method. For detailed guidelines with respect to the choice of an appropriate reference method and the proper interpretation of the results, see Geiser, Eid, and Nussbeck (2008).

We can see that the CT-C($M-1$) model in Figure 1 is for a design with two traits and three methods. In Figure 1, Method 1 serves as reference and thus does not include method factors. The model contains four trait-specific method factors M_{tm}; two for Method 2 and two for Method 3. Standardized trait factor loadings (indicated as λ in the present article) of the nonreference indicators quantify the degree of convergent validity relative to the reference method. Standardized method factor loadings (indicated δ in the present article) measure the degree of method specificity (discrepancy between a given nonreference method and the reference method). Trait factor correlations indicate discriminant validity with respect to the reference method. Method factor correlations within the same method

indicate the degree of generality of method effects across traits. Method factor correlations across different methods indicate to which extent non-reference methods share a common perspective above and beyond what each method shares with the reference method. Trait factors are not allowed to correlate with method factors pertaining to the same trait because method factors are defined as residuals with respect to the trait factors (Eid, 2000). Associations between the reference method and the non-reference methods are captured by the standardized trait factor loadings of the non-reference indicators on the reference trait factors.

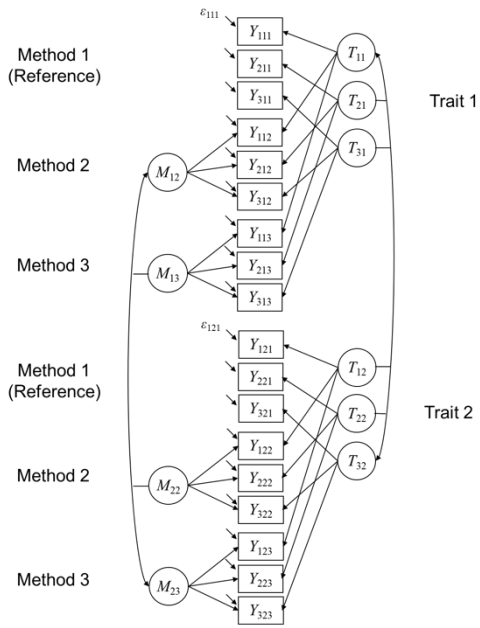

Figure 2:
Multiple-indicator CT-C($M-1$) model with indicator-specific trait factors for a 2 traits × 3 methods design and three indicators per trait-method combination.

In the context of multi-rater studies, different rater types often use the same or similarly worded questionnaires or items. For example, in a multi-rater study on depression, a child self-report item may be worded as "I often feel depressed" with the corresponding parent-report item being "My child often feels depressed." As a result of identical or at least similar item wording across rater types, item- or indicator-specific effects might carry over across raters resulting in the same items being more highly correlated across raters than items differing in content. To take indicator-specific effects into account, Eid et al. (2008) presented an extended version of the multiple-indicator CT-C($M-1$) model that includes indicator-specific trait factors (see Figure 2). In this model, each indicator has its own trait factor, and the indicator-specific trait factors can all be correlated. Of importance, trait factors pertaining to the same trait, but different indicators, do not have to be perfectly

correlated as is implied by models with general trait factors such as the model depicted in Figure 1. Nonetheless, correlations between indicator-specific traits are often substantial.

Previous Studies Examining the Performance of Multiple-Indicator CFA-MTMM Models

Even though multiple-indicator models are more flexible and imply more realistic assumptions about method effects than do single-indicator models, relatively little is known about their applicability and performance across a wide range of conditions. In addition, there appear to be relatively few applications of multiple-indicator models to date. In our systematic review (Geiser & Simmons, in press), we were only able to identify 20 applications with correlated trait-specific method factors most of which were reported as part of methodological papers. Perhaps applied researchers are unsure about these models given how little is known about their performance. In addition, it is well known from various applications and simulation studies that single-indicator CFA-MTMM models can be prone to estimation problems such as nonconvergence and improper parameter estimates (Eid, 2000; Kenny & Kashy, 1992; Marsh, 1989). Multiple-indicator models are more complex and include more parameters than single-indicator models. Given the frequent occurrence of estimation problems with single-indicator CFA-MTMM models, it is important to gain more insights into the performance of multiple-indicator CFA-MTMM models as well, which appear to be promising alternatives to single-indicator models. To date, only few studies have examined the performance of multiple-indicator CFA-MTMM models under a limited set of conditions.

Nussbeck, Eid, and Lischetzke (2006) studied the performance of the multiple-indicator CT-C($M-1$) model for ordinal indicators using mean and variance adjusted weighted least squares estimation (WLSMV). Geiser (2009) examined an extension of the multiple-indicator CT-C($M-1$) model for longitudinal data using continuous indicators and maximum likelihood estimation. Both these studies found minimal parameter and standard error bias, but revealed issues with the accuracy of the chi-square test of model fit when simulating correctly specified multiple-indicator CT-C($M-1$) models. In Nussbeck et al.'s (2006) study, WLSMV chi-square tests of model fit were relatively accurate, but tended to show Type-I error rates that were somewhat lower than the nominal level of alpha (e.g., .01 rather than .05). In contrast, in Geiser (2009), it was found that maximum likelihood chi-square values were inflated and led to rejection of too many correctly specified models even in relatively large samples. Both studies examined only correctly specified models.

The finding of an inflated chi-square statistic is well in line with the "model-size effect" that has consistently been found in studies of CFA models with many observed variables (indicators; Herzog, Boomsma, & Reinecke, 2007; Kenny & McCoach, 2003; Moshagen, 2012; Shi, Lee, & Terry, 2018). These studies all revealed that as the number of observed variables included in a CFA model increases, the accuracy of the empirical chi-square approximation decreases in such a way that Type-I error rates increase (too many correctly specified models are rejected). This effect is particularly severe in samples of small to moderate size.

Multiple-indicator CFA-MTMM models by design use many observed variables. For example, 27 indicators are used in the typical 3 Traits × 3 Methods design with three indicators Y_{itm} per trait-method unit. The multiple-indicator CT-C(M – 1) model with general trait factors for this design (see Figure 1) has 288 degrees of freedom. When more traits, methods, or indicators are used, the number of observed variables increases further. Therefore, when applying multiple-indicator CFA-MTMM models, researchers must expect the model-size effect and run the risk of incorrectly rejecting correctly specified models too frequently unless they use a very large sample. This effect can cause researchers to erroneously abandon adequate models and/or to resort to unnecessarily highly parameterized alternative models that seemingly show a better fit. It is therefore important to study the performance of available correction procedures for the chi-square statistic that we discuss later in this article.

The two previous studies examining multiple-indicator CFA-MTMM models that we know of (Geiser, 2009; Nussbeck et al., 2006) are limited in several ways. First, these studies included a rather limited set of conditions. Both studies examined only one set of parameter estimates that was derived from a single real data application, respectively. Second, these studies did not examine the performance of correction procedures for the model-size effect to find out whether more accurate chi-square tests of model fit could be obtained through such procedures. Third, previous studies only examined model versions with general trait factors (Figure 1) that are frequently misspecified in actual empirical applications due to the presence of indicator-specific effects (i.e., less than perfectly unidimensional indicators). In practice, the more complex model version with indicator-specific trait factors (Figure 2) is often more adequate. Models with indicator-specific traits are even more complex in terms of the number and types of free parameters. To our knowledge, the performance of models with indicator-specific trait factors has not yet been examined in simulation research. Fourth, previous studies only examined correctly specified models. Therefore, to date, nothing is known about the consequences of misspecification, including the question of whether the (corrected and uncorrected) chi-square statistics have sufficient statistical power to reveal such misspecifications. The present simulation study was designed to shed more light on these questions. Below, we first discuss chi-square correction procedures for the model-size effect that have been proposed in the literature. Subsequently, we present the results of our simulation study of two different versions of the multiple-indicator CT-C(M – 1) model.

Correction Procedures for the Model-Size Effect

When the number of variables in a confirmatory factor or structural equation model is large and the sample size is relatively small, the maximum likelihood chi-square statistic tends to over-reject correctly specified models (i.e., there is a Type-I error inflation in these cases; see discussion above). Several correction procedures have been proposed to adjust the chi-square statistic in these cases. These correction procedures are discussed in detail in Yuan, Tian, and Yanagihara (2015). The goal of all model-size correction procedures is to shrink the empirical chi-square statistic so that the corrected statistic will result in more accurate Type-I error rates that do not lead to an over-rejection of correctly specified models. Yuan et al. (2015) compared different procedures that are based on analytical solutions to an empirical correction procedure that these authors developed based on simulations. In Yuan et al.'s (2015) procedure, the empirical maximum likelihood chi square test statistic T_{ML} is corrected by multiplying T_{ML} with a correction factor e such that the corrected statistic $T_{MLe} = eT_{ML}$, where $e = [N - (2.381 + 0.361p + 0.006q)]/(N - 1)$, N indicates the sample size, p indicates the number of observed variables, and q indicates the number of free parameters estimated in the model. The larger p, the greater the shrinkage of the T_{ML} statistic according to this correction. In addition, the number of free parameters (q) also has a small impact on the chi-square statistic that is also considered through the correction factor e.

Yuan et al. (2015) found that their empirical procedure outperformed previously described analytical correction procedures. A study by Shi et al. (2018) confirmed that the Yuan et al. (2015) empirical correction procedure performed best for simple-structure CFA models with a homogeneous loading pattern of the indicators (all indicators had equal loadings and low reliabilities of .49). In this article, we only studied the Yuan et al. (2015) correction given that it has been shown to outperform all previous procedures.

Even though the Yuan et al. (2015) correction procedure appears to show good performance for conventional (simple-structure) CFA models, this procedure to our knowledge has not yet been tested for multiple-indicator CFA-MTMM models. Compared to conventional simple-structure CFA models, multiple-indicator CFA-MTMM models are special in that they include cross-loadings of indicators on multiple factors and many trait and method factor correlations. Moreover, in previous studies of simple-structure CFA models, indicators were simulated with homogeneous loadings and low reliabilities (standardized loading of .7 which corresponds to a reliability of .49). In MTMM studies, a homogeneous loading pattern of all indicators is not expected. For example, in the CT-C($M - 1$) approach, non-reference indicators typically have smaller loadings on the reference trait factor as compared to reference indicators due to the presence of method effects which result in less than perfect convergent validity. Another reason for heterogeneous loadings can be differences in the reliabilities (i.e., the amount of random error variance) between indicators either within the same method or across different methods.

Goals of the Present Study

In summary, multiple-indicator CFA-MTMM models represent particularly complex CFA models as they contain many indicators, cross-loadings, loadings that likely differ in size at least across methods, potentially indicator-specific trait factors, and many factor correlation parameters to be estimated. To date, limited simulation work is available regarding the performance of multiple-indicator CFA-MTMM models. We know of no simulation studies that have examined the applicability and accuracy of correction procedures for the chi-square test of model fit for these models or simulations of multiple-indicator CFA-MTMM models with indicator-specific traits. The purpose of the present study was to (1) study the performance (convergence rates, improper parameter estimates, parameter and standard error bias) of both the multiple-indicator CT-C(*M* – 1) model with general and indicator-specific factors across a range of realistic conditions and (2) examine the performance of Yuan et al.'s (2015) chi-square correction procedure that accounts for the model-size effect. To this end, we studied correctly and incorrectly specified models so that we could determine both Type-I error rates (rates of incorrectly rejecting proper models) and statistical power (correct rejection of misspecified models).

Method

Using Mplus 8 (Muthén & Muthén, 1998-2017), we simulated normally distributed data based on multiple-indicator CT-C(*M* – 1) population models (see Figures 1 and 2) each with three continuous indicators Y_{1tm}, Y_{2tm}, and Y_{3tm} per trait-method combination.[4] The choice of our simulation conditions was informed by our previous literature review of applied multiple-indicator CFA-MTMM models (Geiser & Simmons, in press). We chose the multiple-indicator CT-C(*M* – 1) approach because to date it is the most frequently used multiple-indicator CFA-MTMM approach (Geiser & Simmons, in press). We simulated a 3 traits × 3 methods design, as this design was the most common design found in our literature review. We assumed continuous observed variables as most multiple-indicator applications to date have used scale-level rather than item-level data. Samples sizes were chosen according to common sample sizes found in our previous literature review. We varied the reliability of the indicators as it is known that model fit statistics are influenced by the level of reliability (e.g., McNeish & Wolf, 2020).

In summary, our simulation design included four fully crossed population conditions:

- Multiple-indicator CT-C(*M* – 1) model version (1 = model with general traits [Figure 1], 2 = model with indicator-specific traits [Figure 2])
- Sample size (1 = small [*N* = 200], 2 = medium [*N* = 500], 3 = large [*N* = 800])

[4]Indicators within each trait-method combination were simulated as tau-parallel in the sense of classical test theory. That is, they had equal factor loadings, error variances, and reliabilities in the population models. Simulating the indicators as parallel simplified our subsequent computations, analyses, and presentation of the results. In addition, no mean structure was included in the simulations.

- Reliability of indicators (1 = low [.50], 2 = high [.81])
- Generalization of method effects across traits (i.e., size of heterotrait-monomethod method factor correlation ϕ; 1 = 1.0, 2 = .9, 3 = .8, 4 = .5, 5 = .3, 6 = 0)

The simulation design thus had 2 (model type) × 3 (sample size) × 2 (reliability) × 6 (method generalization) = 72 cells. All observed variables were simulated to be in *z*-score metric in the population models. We varied convergent validity across non-reference methods within each cell of the design such that some indicators pertaining to Method 2 showed high convergent validity, and indicators pertaining to Method 3 showed low convergent validity. Likewise, discriminant validities (trait factor correlations) were varied within each condition. We studied two indicator reliability conditions because applied studies often use item parcels that may consist of a limited number of items and thus show rather low reliabilities. A detailed description of the specific parameter values used in the simulation as well as all Mplus input and output files can be found in the online supplemental materials at https://osf.io/4hgyt/.

In the first part of our simulations, we generated 1,000 data sets (replications) for each cell of the design and fit correctly-specified models to the data using maximum likelihood estimation. We examined convergence rates, improper solutions, parameter bias, standard error bias, the model-size effect on chi-square, and the performance of Yuan et al.'s (2015) chi-square correction procedure for correctly-specified models. In the second part of our simulation study, we examined the consequences of erroneously fitting a CT-C(M – 1) model with general trait factors (Figure 1) to data generated from the indicator-specific traits condition (Figure 2). We were interested in determining whether Yuan et al.'s (2015) chi-square correction procedure would provide adequate statistical power to detect this misspecification and whether parameters and standard errors for the most relevant parameters were biased when fitting a model with general traits to data generated from indicator-specific traits.

Results

Correctly Specified Models

Convergence Rates. For both model versions (general and indicator-specific traits), the CT-C(M – 1) model showed 100% convergence in all conditions except (1) the three conditions with low sample size (N = 200), low reliability (.5), and small (ϕ = .5 and ϕ = .3) or zero method factor correlations. For these conditions, convergence rates were still ≥ 98.1%. In addition, a single replication of the indicator-specific traits model simulated for a medium sample size (N = 500), high reliability (.81), and small method correlations (ϕ = .3) also did not converge, reducing the convergence rate for this condition to 99.9%.

Improper Solutions and Non-Positive Definite Matrices. Negative error variance estimates occurred only in the three above-mentioned small-sample, low-reliability conditions that also showed some convergence issues. In these conditions, the rate of improper error

variance estimates was still low (≤ 3.7%). Non-positive definite latent variable covariance matrices were more common. This can be a sign of improper latent variance or correlation estimates or of linear dependencies among latent variables. Population models with high correlations among trait or method factors are more susceptible to non-positive definite matrices. For the model version with general trait factors, the rates were zero or very low (≤ 4.7%) in all except the following conditions:

- rates were ≥ 98.2% when the population method factor correlations were specified to be very high (ϕ = .9), but not perfect (ϕ = 1).[5]
- for method factor correlations of ϕ = .8, the rates were low (0% to 9.7%) in the high reliability conditions, but rather high (31.8% to 77.5%) in the low reliability conditions.

For the model with indicator-specific trait factors, the rates of non-positive definite latent matrices were low (≤ 6.3%) when the sample size was moderate (500) or large (800) in conjunction with strong indicator reliabilities (.81) and method factor correlations that were either perfect (ϕ = 1) or ≤ .8. In all other conditions, the rates ranged between 33.6% and 100%.

Model Fit. Tables 1 provides a summary of model fit statistics for each condition of our simulations of the correctly specified general traits model. Results were very similar for the indicator-specific traits model. Therefore, results for this model are not shown here but are included in the supplemental materials. The expected model-size effect occurred for both model types. Average uncorrected chi-square values were larger than theoretically expected for correctly specified models. As a result, chi-square rejection (Type-I error) rates for a nominal alpha level of .05 were inflated in all conditions, leading to rejection of too many correctly specified models. The indicator-specific traits model showed slightly more accurate rejection rates than did the model with general traits; nonetheless, chi-square values were inflated in all conditions.

Overall, in the small sample size condition (N = 200), Type-I error rates were inflated to between 20% and 23.4% (general trait model) and to between 17.6% and 22.4% (indicator-specific traits) for a nominal alpha level of 5%. For N = 500, Type-I error rates were reduced to between 9.7% and 11.2% (general trait model) and to between 8.8% and 10.1% (indicator-specific trait model). In the largest sample size condition (N = 800), Type-I error rates were relatively accurate (between 6.8% and 8.1% for general traits and between 6.7% and 7.4% for indicator-specific traits), but still slightly inflated.

[5]These problems likely did not occur in the perfect-correlation (ϕ = 1.0) conditions because in these conditions, the method factors in the population model were combined into a single general method factor as is implied by ϕ = 1.0.

Table 1:

Model Fit Statistics for the Correctly Specified Multiple-Indicator CT-C(M – 1) Model with General Traits Across Simulation Conditions

Condition code	χ^2 *M*	χ^2 *SD*	*df*	χ^2 rejection rate (alpha = .05)	Yuan et al. (2015) corrected χ^2 rejection rate (alpha = .05)	RMSEA *M*	RMSEA *SD*	SRMR *M*	SRMR *SD*
111	323.323	27.355	302	.230	.063	0.017	0.011	0.050	0.005
112	307.822	26.997	288	.219	.056	0.016	0.012	0.048	0.004
113	307.683	26.985	288	.220	.057	0.016	0.012	0.048	0.004
114	307.137	27.004	288	.213	.060	0.016	0.012	0.048	0.004
115	306.771	27.036	288	.208	.059	0.016	0.012	0.047	0.003
116	306.685	27.102	288	.200	.058	0.016	0.012	0.047	0.003
121	323.523	27.379	302	.234	.068	0.017	0.011	0.048	0.012
122	308.423	26.992	288	.231	.065	0.017	0.012	0.045	0.011
123	308.451	26.995	288	.228	.068	0.017	0.012	0.044	0.010
124	308.318	27.053	288	.222	.070	0.017	0.012	0.041	0.007
125	308.197	27.114	288	.221	.078	0.017	0.012	0.040	0.006
126	308.040	27.185	288	.216	.074	0.017	0.012	0.039	0.005
211	309.662	26.300	302	.106	.050	0.007	0.007	0.032	0.003
212	295.238	25.982	288	.111	.062	0.007	0.007	0.030	0.003
213	295.204	25.990	288	.112	.062	0.007	0.007	0.030	0.003
214	295.028	25.984	288	.109	.055	0.007	0.007	0.030	0.002
215	294.888	25.923	288	.100	.054	0.007	0.007	0.030	0.002
216	294.806	25.751	288	.097	.053	0.007	0.007	0.030	0.002
221	309.785	26.336	302	.103	.051	0.007	0.007	0.030	0.008
222	295.396	26.101	288	.107	.063	0.007	0.007	0.029	0.007
223	295.467	26.088	288	.109	.062	0.007	0.007	0.028	0.006
224	295.419	26.034	288	.109	.059	0.007	0.007	0.026	0.005
225	295.318	25.962	288	.102	.057	0.007	0.007	0.025	0.004
226	295.150	25.387	288	.104	.061	0.007	0.007	0.025	0.003
311	305.911	25.216	302	.074	.051	0.005	0.005	0.025	0.002
312	291.763	24.945	288	.080	.057	0.005	0.005	0.024	0.002
313	291.763	24.960	288	.078	.057	0.005	0.005	0.024	0.002
314	291.707	24.934	288	.073	.055	0.005	0.005	0.024	0.002
315	291.665	24.862	288	.069	.054	0.005	0.005	0.024	0.002
316	291.644	24.727	288	.070	.052	0.005	0.005	0.024	0.001
321	305.827	25.294	302	.068	.050	0.005	0.005	0.024	0.006
322	291.759	25.122	288	.080	.056	0.005	0.005	0.023	0.005
323	291.802	25.155	288	.081	.058	0.005	0.005	0.022	0.005
324	291.823	25.051	288	.079	.058	0.005	0.005	0.020	0.004
325	291.833	24.928	288	.076	.058	0.005	0.005	0.020	0.003
326	291.836	24.720	288	.069	.054	0.005	0.005	0.019	0.003

Note. RMSEA = root mean square error of approximation. SRMR = standardized root mean square residual. The first digit in the condition code is for sample size (1 = 200; 2 = 500; 3 = 800). The second digit is for reliability (1 = .5; 2 = .81). The third digit is for population method factor correlation (1 = 1; 2 = .9; 3 = .8; 4 = .5; 5 = .3; 6 = 0). The χ^2 rejection rate (alpha = .05) columns give an estimate of the Type-I error rate for each condition. Models in Conditions *xx*1 have more *df* because they include fewer method factors (method factors for the same method are perfectly correlated and thus collapses into one factor).

Application of Yuan et al.'s (2015) chi-square correction procedure led to Type-I error rates that were closer to the nominal alpha level of .05 across all conditions. There was still a slight inflation of Type-I error rates in most conditions even after correction; however, the Yuan et al. (2015) corrected Type-I error rates were much closer to .05 than were the uncorrected rates. The "worst" condition showed a corrected rejection rate of .078 (as opposed to .221 uncorrected), which still represents a rather large improvement. In summary, Yuan et al.'s (2015) correction procedure appeared to work well for both the general and the indicator-specific trait versions of the multiple-indicator CT-C(M – 1) approach. For both model types, regular (uncorrected) indices of approximate fit such as the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR; see Tables 1 and 2) would not have led to over-rejection of the correctly specified population models according to common standards (e.g., Hu & Bentler, 1999), so that we did not study Yuan et al.'s (2015) correction for these statistics.

Parameter Estimate Bias. We calculated percent parameter estimate bias (%peb) as

$$\%\text{peb} = 100 \cdot \frac{E(\hat{P}) - P}{P},$$

where $E(\hat{P})$ indicates the average parameter estimate across replications, and P indicates the true population parameter. The %peb measure allows us to determine the percent bias in parameter estimates relative to the true population parameter for each simulation condition.

We averaged the absolute values of %peb across parameters within the same parameter type. For example, we averaged |%peb| values across the reference trait factor loadings pertaining to the reference method which all had the same true population values within a given condition. We used absolute values of peb to avoid misleading effects of positive and negative bias values potentially averaging out for a given parameter type.

The averaged |%peb| values were close to zero for all parameter types under all conditions. The highest |%peb| values found were ≤ 3.1% for error variance parameters in the small sample, low reliability conditions. The exact |%peb| values for all parameter types and both model versions can be found in the online supplemental materials.

Standard Error Bias. We calculated percent standard error bias (%seb) as

$$\%\text{seb} = 100 \cdot \frac{E(SE_p) - SD_p}{SD_p},$$

where SE_p indicates the average standard error estimate for Parameter P across replications, and SD_p indicates the standard deviation of the parameter estimate for Parameter P across replications.

The %seb measure allows us to determine the percent bias in standard error estimates by comparing the estimated large-sample standard errors to the standard deviation of the simulated sampling distribution for each parameter.

Average percent absolute standard error bias across simulation conditions is shown in Table 2 for the indicator-specific traits model. Results were very similar for the general traits model (although bias values were slightly lower for that model), which is why we include only the |%seb| values for the indicator-specific model here. The corresponding table for the model with general trait factors can be found in the supplemental materials.

Standard error bias was generally small for all parameter types (i.e., below 5% for most conditions) for both models. However, standard errors for specific parameters in some of the small sample (N = 200), low reliability (.5) conditions were negatively biased (estimated to be too small; see bold-face entries in Table 2). In these conditions, the standard errors associated with the smaller method factor loadings (δ_{it2}), the measurement error variances, and the method factor correlations were underestimated by up to 21.8% in the conditions of zero, small (.30), or moderate (.50) method factor correlations across traits. Standard errors for the larger method factor loadings (δ_{it3}) and other parameters showed relatively small bias below 10% in these conditions.

Misspecified Models

Convergence Rates. When (incorrectly) fitting a model version with general traits to data generated from the indicator-specific traits model, convergence rates were still ≥ 95.7% in all conditions.

Improper Solutions and Non-Positive Definite Matrices. The rate of improper error variance estimates in the misspecified simulation was ≤ 5.1% across all conditions. Non-positive definite latent variable covariance matrices were uncommon except (1) in the low sample size, low reliability conditions (rates up to 99.9%) and (2) in other conditions when the method factor correlations in the population model were set at .8 or .9 (rates between 71.9% and 100%). All other conditions showed rates ≤ 0.1%.

Model Fit. Table 3 provides a summary of model fit statistics for each condition in the misspecified model simulation. Recall that in this case, fit statistics should indicate misspecification and lead to rejection of the models. Chi-square rejection rates here indicate power to detect misspecified models. Table 3 shows that uncorrected chi-square rejection rates (estimated power) ranged between 70% and 100%. Power was lowest in the small sample, low reliability conditions. Yuan et al.'s (2015) corrected chi-square rejection rates were generally very similar, except in the small sample, low reliability conditions. In these conditions, power was reduced from between 70% and 74.1% to between 43.8% and 50.4%.

Table 2:

Percent Standard Error Bias in the Correctly Specified Multiple-Indicator CT-C(M – 1) Model with Indicator-Specific Traits Across Simulation Conditions

	N = 200						*N = 500*						*N = 800*					
Parameter	ϕ=*1*	ϕ=*.9*	ϕ=*.8*	ϕ=*.5*	ϕ=*.3*	ϕ=*0*	ϕ=*1*	ϕ=*.9*	ϕ=*.8*	ϕ=*.5*	ϕ=*.3*	ϕ=*0*	ϕ=*1*	ϕ=*.9*	ϕ=*.8*	ϕ=*.5*	ϕ=*.3*	ϕ=*0*
Low reliability [$Rel(Y_{tm})$ = .50]																		
λ_{it1}	4.0	4.6	4.6	4.8	4.9	4.8	3.1	2.8	2.9	3.1	3.1	3.2	2.7	2.7	2.6	2.6	2.6	2.4
λ_{it2}	5.4	6.5	6.4	6.4	6.3	6.4	2.8	2.6	2.6	2.7	2.7	2.8	2.4	2.5	2.5	2.6	2.5	2.4
λ_{it3}	5.6	7.1	7.2	7.4	7.5	7.6	2.3	2.8	2.9	2.9	2.9	2.8	1.7	1.6	1.6	1.4	1.3	1.2
δ_{it2}	4.3	8.2	9.2	**15.4**	**20.5**	**21.8**	1.8	2.8	3.2	5.0	6.6	6.3	1.7	2.5	2.7	3.4	4.0	3.8
δ_{it3}	2.8	4.6	5.1	6.6	7.5	9.7	1.7	2.4	2.6	3.2	3.3	3.3	1.9	1.9	1.9	2.0	2.2	2.2
$Var(\varepsilon_{itm})$	4.2	5.5	5.8	9.1	**15.3**	**15.4**	1.8	2.0	2.0	2.4	3.0	3.2	1.3	1.3	1.3	1.3	1.4	1.8
ϕ_T	0.0	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ϕ_M	3.3	7.4	7.8	9.9	**11.7**	**11.7**	1.1	2.2	2.3	2.5	2.8	3.9	2.2	2.3	2.2	2.3	2.5	2.6
High reliability [$Rel(Y_{tm})$ = .81]																		
λ_{it1}	1.6	1.4	1.4	1.5	1.4	1.6	1.5	1.3	1.3	1.2	1.1	1.1	2.3	2.0	2.0	1.9	2.0	1.9
λ_{it2}	3.8	5.0	5.1	5.0	4.8	4.4	2.0	2.0	2.1	2.2	2.2	2.2	2.0	2.1	2.4	2.8	2.9	2.8
λ_{it3}	3.2	4.3	4.3	4.5	4.7	5.3	1.9	1.3	1.4	1.8	1.8	1.7	1.5	1.1	1.4	1.6	1.3	0.8
δ_{it2}	2.7	3.3	3.3	3.7	4.0	4.0	1.3	1.2	1.4	1.6	1.7	1.8	1.3	1.4	1.3	1.2	1.4	2.0
δ_{it3}	2.2	2.8	2.6	2.4	2.0	2.1	1.7	2.2	2.3	2.5	2.4	1.8	1.2	1.6	1.7	1.3	1.7	1.9
$Var(\varepsilon_{itm})$	3.1	4.1	4.2	4.1	4.1	4.4	1.9	1.9	1.8	1.9	1.9	1.9	1.6	1.3	1.2	1.4	1.4	1.5
ϕ_T	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ϕ_M	2.1	3.2	3.3	3.6	3.6	3.3	1.6	1.5	1.4	1.4	1.5	2.0	2.7	1.8	1.6	1.7	2.0	1.8

Note. ϕ = latent correlation between method factors pertaining to the same method but different traits in the population models. $Rel(Y_{tm})$ = reliability. λ_{itm} = trait factor loading (i = indicator, t = trait, m = method; reference method: m = 1; non-reference methods: m = 2, 3). δ_{itm} = method factor loading. $Var(\varepsilon_{itm})$ = measurement error variance. ϕ_T = correlation between trait factors. ϕ_M = correlation between method factors. Table entries show average absolute bias values in percent. Bias values for loading and error variance parameter estimates were averaged across the three indicators within each method as these indicators had the same parameters in the population model. Bold face entries indicate *SE* bias > |10%|.

Indices of approximate model fit varied in their ability to detect the misspecification. According to Hu and Bentler (1999), satisfactory approximate fit is indicated (roughly) by RMSEA ≤ .06, comparative fit index (CFI) and Tucker-Lewis index (TLI) ≥ .95, and SRMR ≤ .08.[6] When applying these criteria, average RMSEA values consistently indicated misfit in the high reliability conditions (values ≥ .071), but (erroneously) indicated decent fit when reliabilities were low (values ≤ .031). A very similar pattern emerged for CFI and TLI. Average SRMR values (erroneously) indicated decent model fit in all conditions (values ≤ .06) and thus would typically not have led to detection of misspecified models regardless of sample size, reliability, and magnitude of method factor correlation.

Parameter Estimate Bias. The averaged |%peb| values for all parameter types are given in the supplemental materials. Bias was generally low (≤ 7.4%) except for estimates of measurement error variances, which were strongly positively biased (overestimated by up to 33.5%) in the high reliability conditions. This means that reliability coefficients in these conditions would be underestimated.

Discussion

The use of multiple-indicator CFA-MTMM models has been recommended by methodologists to overcome overly restrictive assumptions of single-indicator CFA-MTMM models (Eid et al., 2003; 2008; Geiser & Simmons, in press; Marsh & Hocevar, 1988). At the same time, little is known so far about the performance of complex CFA-MTMM models under different conditions. Studying the adequacy of such models is important given that even the less complex single-indicator CFA-MTMM models can be prone to estimation problems. In the current study, we examined two versions of the multiple-indicator CT-C(M – 1) approach, one of which had not previously been studied through simulation work. Our findings confirm results of previous research (Geiser, 2009; Nussbeck et al., 2006), according to which the multiple-indicator CT-C(M – 1) approach shows high convergence rates and negligible parameter bias even in samples as small as N = 200. We also found that improper (i.e., negative) error variance estimates were uncommon with this approach, even in small samples.

However, non-positive definite latent variable covariance matrices occurred frequently in our simulations. This is likely because we simulated models in which correlations between latent variables were very high (.8 or .9) in some conditions. As a result, sampling error can lead to correlation estimates > 1.0 in simulated data sets. In addition, high correlations among latent variables can lead to linear dependencies. In that case, improper parameter estimates may not occur, but a warning message about a non-positive definite latent variable covariance matrix can still be issued by statistical software programs. In practice, researchers encountering such messages in applications of the CT-C(M – 1) approach

[6]The universal application Hu and Bentler's (1999) guidelines for approximate fit has been criticized as these guidelines are based on a limited set of simulation conditions. McNeish and Wolf (2020) showed that dynamic fit index cutoffs can lead to more valid results. We nonetheless used Hu and Bentler's (1999) guidelines in the present work given that the validity of dynamic fit index cutoffs has not yet been demonstrated for CFA-MTMM models with a complex factor and loading structure (McNeish & Wolf, 2020).

should carefully study all latent variable parameter estimates. In cases of perfect or near-perfect correlations between latent variables, these latent variables may be collapsed into a single latent variable.

Although standard errors were generally adequate for correctly specified models, we found that specific parameters showed underestimated standard errors when the sample size was small (N = 200), indicator reliabilities were low (.5), and method factor correlations indicating the degree of generalization of method effects across traits were zero, small, or moderate in size. Affected parameters in these conditions were the smaller method factor loadings, the measurement error variances, and method factor correlations, for which the standard errors were underestimated by up to 21.8%. Hence, in these conditions, the precision of estimation of the parameters in question would have been overestimated based on the estimated standard errors (i.e., estimated confidence intervals would be too narrow).

In addition, the underestimated standard errors could lead to Type-I error inflation when testing method factor loadings and method factor correlations for statistical significance. As a result, a researcher may incorrectly claim that method effects are present when measures in fact do not show method effects. In addition, spurious associations between different methods may be over-interpreted. The former Type-I error (incorrectly assuming that method effects are present when they are absent) is probably less of a concern for actual substantive applications as method effects in the social sciences are typically present and rather strong. We found no substantial *SE* bias for the indicators that showed stronger method loadings (which may be more realistic) in the above-mentioned conditions. Incorrectly assuming that method effects are present (when in fact measures show perfect convergent validity) should be a rare problem in the real world.

Potential Type-I errors regarding between-method associations may be more problematic. For example, one might incorrectly claim that parent and teacher reports show a common perspective on child problem behaviors when in fact this may not be true. This incorrect claim could have real consequences. For instance, based on this finding, one might decide to collect only parent reports (but not teacher reports) due to ostensible redundancies. In that case, important information about differences in rater perspectives may be missed by the investigator.

To avoid these issues, we recommend that researchers using multiple-indicator CFA-MTMM models aim for samples larger than N = 200 or choose highly reliable indicators. In cases in which this is not possible, researchers might use a more stringent alpha level for tests of statistical significance (e.g., .01 instead of .05) or employ robust standard errors.

Table 3:
Model Fit Statistics for the Misspecified Multiple-Indicator CT-C(M – 1) Model with General Traits Across Simulation Conditions

Condition code	χ^2 *M*	χ^2 *SD*	*df*	χ^2 rejection rate (alpha = .05)	Yuan et al. (2015) corrected χ^2 rejection rate (alpha = .05)	RMSEA *M*	RMSEA *SD*	CFI *M*	CFI *SD*	TLI *M*	TLI *SD*	SRMR *M*	SRMR *SD*
111	365.128	31.161	302	0.741	0.504	0.031	0.009	0.968	0.015	0.963	0.018	0.053	0.004
112	347.385	30.587	288	0.719	0.490	0.031	0.010	0.969	0.015	0.962	0.019	0.052	0.004
113	346.915	30.564	288	0.721	0.482	0.031	0.010	0.968	0.016	0.961	0.020	0.051	0.004
114	345.345	30.454	288	0.709	0.458	0.030	0.010	0.966	0.017	0.959	0.021	0.051	0.004
115	344.340	30.470	288	0.706	0.447	0.030	0.010	0.966	0.018	0.959	0.022	0.051	0.003
116	344.051	30.407	288	0.700	0.438	0.030	0.010	0.966	0.018	0.958	0.022	0.051	0.003
121	639.259	51.397	302	1.000	1.000	0.075	0.006	0.939	0.009	0.929	0.011	0.058	0.010
122	600.223	49.483	288	1.000	1.000	0.073	0.006	0.941	0.009	0.928	0.011	0.060	0.011
123	597.719	49.209	288	1.000	1.000	0.073	0.006	0.938	0.010	0.925	0.012	0.058	0.010
124	593.519	48.845	288	1.000	1.000	0.073	0.006	0.934	0.011	0.920	0.013	0.054	0.007
125	591.888	48.731	288	1.000	1.000	0.072	0.006	0.933	0.011	0.918	0.013	0.053	0.006
126	590.921	48.713	288	1.000	1.000	0.072	0.006	0.933	0.011	0.918	0.013	0.052	0.005
211	415.983	34.235	302	0.983	0.968	0.027	0.004	0.977	0.007	0.973	0.008	0.037	0.003
212	396.596	33.724	288	0.979	0.963	0.027	0.004	0.977	0.007	0.972	0.009	0.036	0.003
213	395.832	33.666	288	0.979	0.961	0.027	0.005	0.976	0.007	0.971	0.009	0.036	0.003
214	393.811	33.497	288	0.973	0.960	0.027	0.005	0.975	0.008	0.969	0.010	0.036	0.002
215	392.594	33.365	288	0.974	0.957	0.027	0.005	0.975	0.008	0.969	0.010	0.036	0.002
216	391.373	33.192	288	0.975	0.953	0.026	0.005	0.974	0.008	0.969	0.010	0.035	0.002
221	1102.466	73.761	302	1.000	1.000	0.073	0.003	0.941	0.006	0.932	0.006	0.045	0.006
222	1029.844	70.201	288	1.000	1.000	0.072	0.003	0.944	0.005	0.931	0.007	0.049	0.008
223	1023.722	69.752	288	1.000	1.000	0.071	0.003	0.941	0.006	0.928	0.007	0.048	0.007
224	1013.557	68.857	288	1.000	1.000	0.071	0.003	0.937	0.006	0.923	0.007	0.044	0.005
225	1009.676	68.409	288	1.000	1.000	0.071	0.003	0.936	0.006	0.922	0.008	0.043	0.004
226	1007.438	68.113	288	1.000	1.000	0.071	0.003	0.936	0.006	0.922	0.008	0.042	0.003
311	477.173	37.427	302	1.000	1.000	0.027	0.003	0.978	0.005	0.974	0.006	0.032	0.002
312	455.335	36.736	288	1.000	1.000	0.027	0.003	0.978	0.005	0.973	0.006	0.031	0.002
313	454.177	36.664	288	1.000	1.000	0.027	0.003	0.977	0.005	0.972	0.006	0.031	0.002
314	451.242	36.417	288	1.000	1.000	0.026	0.003	0.976	0.005	0.970	0.007	0.031	0.002
315	449.558	36.197	288	1.000	0.999	0.026	0.003	0.975	0.006	0.970	0.007	0.031	0.002
316	447.745	35.917	288	1.000	0.998	0.026	0.003	0.975	0.006	0.970	0.007	0.030	0.002
321	1577.428	87.740	302	1.000	1.000	0.073	0.002	0.942	0.004	0.932	0.005	0.041	0.004

Condition code	χ^2 *M*	χ^2 *SD*	*df*	χ^2 rejection rate (alpha = .05)	Yuan et al. (2015) corrected χ^2 rejection rate (alpha = .05)	RMSEA *M*	RMSEA *SD*	CFI *M*	CFI *SD*	TLI *M*	TLI *SD*	SRMR *M*	SRMR *SD*
322	1470.969	83.116	288	1.000	1.000	0.072	0.003	0.944	0.004	0.932	0.005	0.046	0.006
323	1461.021	82.587	288	1.000	1.000	0.071	0.003	0.941	0.004	0.928	0.005	0.044	0.005
324	1444.931	81.590	288	1.000	1.000	0.071	0.003	0.937	0.005	0.924	0.005	0.041	0.004
325	1439.020	81.185	288	1.000	1.000	0.071	0.002	0.936	0.005	0.922	0.006	0.040	0.003
326	1435.922	80.998	288	1.000	1.000	0.071	0.002	0.936	0.005	0.922	0.006	0.039	0.003

Note. RMSEA = root mean square error of approximation. CFI = comparative fit index. TLI = Tucker-Lewis index. SRMR = standardized root mean square residual. The first digit in the condition code is for sample size (1 = 200; 2 = 500; 3 = 800). The second digit is for reliability (1 = .5; 2 = .81). The third digit is for population method factor correlation (1 = 1; 2 = .9; 3 = .8; 4 = .5; 5 = .3; 6 = 0). The χ^2 rejection rate (alpha = .05) columns give an estimate of power to reject a misspecified model. Models in Conditions *xx*1 have more *df* because they include fewer method factors (method factors for the same method are perfectly correlated and thus collapses into one factor).

In our simulations of correctly specified models, we replicated the model-size effect according to which the maximum likelihood chi-square test statistic tends to be inflated due to the use of many observed variables (27 in the case of a 3 × 3 × 3 multiple-indicator MTMM design). This effect leads to rejection of too many correctly specified models. We found that Yuan et al.'s (2015) empirical correction procedure resulted in chi-square rejection rates that were much closer to the nominal Type-I error level across all conditions compared to uncorrected chi-square rejection rates. We therefore recommend that researchers using multiple-indicator CFA-MTMM models and maximum likelihood estimation employ Yuan et al.'s (2015) correction procedure when testing their models to avoid over-rejection of correctly specified models.

The only issue with this correction in our simulations was that it led to reduced power (between 43.8% and 50.4% in our simulations) to detect misspecified models in small samples ($N = 200$) when indicators had low reliabilities ($Rel = .5$). Under these conditions, the correction procedure may lead to the (erroneous) acceptance of misspecified models in about every other case. To avoid this problem, we recommend that (1) researchers using small samples choose highly reliable indicators and (2) researchers using indicators with rather low reliabilities choose samples larger than $N = 200$ to achieve sufficient power to detect misspecified models when using Yuan et al.'s (2015) correction procedure for multiple-indicator CT-C($M - 1$) models. Likewise, we recommend that researchers interpret indices of approximate fit (RMSEA, CFI, TLI, SRMR) with caution when indicators have low reliabilities, as these indices appear to be less sensitive to misspecifications when indicator reliabilities are low regardless of the sample size.

In a recent paper, McNeish and Wolf (2020) criticized the use of fixed guidelines for indices of approximate fit such as Hu and Bentler's (1999) guidelines. As an alternative, McNeish and Wolf (2020) proposed the use of dynamic cutoff values for fit statistics based on simulations for the specific application at hand. It would be useful to study the appropriateness of their approach for CFA-MTMM models in future research.

In summary, our simulations showed that multiple-indicator CT-C($M - 1$) models perform well unless a small sample (200 or smaller) *and* unreliable indicators are used, in which case several different problems may occur. The use of indicators with higher reliabilities appears to (partly) compensate for the negative effects of a small sample and vice versa.

Researchers frequently apply the multiple-indicator CT-C($M - 1$) model with general trait factors. In practice, indicators may show indicator-specific effects leading to misspecification of the general-traits approach. Our study revealed that this type of misspecification may be difficult to detect based on corrected chi-square statistics and indices of approximate model fit, especially when indicator reliabilities are low. We therefore recommend that researchers using equivalent scales across methods apply both model versions and carefully compare the parameter estimates. In case of substantial discrepancies, we recommend that the indicator-specific model version be chosen as it implies less restrictive assumptions with respect to the unidimensionality of indicators and thus is less likely to result in biased parameter estimates.

Goals for Future Research

In our simulations, we examined only the multiple-indicator CT-C(*M* – 1) approach because to date this approach is most frequently used in multiple-indicator MTMM designs (Geiser & Simmons, in press). In future studies, other types of multiple-indicator CFA-MTMM models (e.g., Pohl & Steyer, 2010; Pohl et al., 2008) should be examined to study the generalizability of our findings to other complex MTMM models. It would also be interesting to compare the performance of different multiple-indicator CFA-MTMM models in future simulations.

In our simulations, we assumed continuous and normally distributed data. Future simulations should also examine consequences of non-normality which occurs frequently in practice. In addition, the model-size effect on chi-square for correctly specified models appears to go in the opposite direction for ordinal indicators and WLSMV estimation (i.e., fewer models than theoretically expected are rejected; Nussbeck et al., 2006). Therefore, it would be interesting to study under which conditions the WLSMV chi-square has sufficient power to detect misspecified models. Finally, we recommend that other types of model misspecification than the one studied here be examined in future work and that the use of dynamic cutoff values for approximate fit statistics (McNeish and Wolf, 2020) be evaluated for CFA-MTMM models.

Disclosure statement.

We have no known conflict of interest to disclose.

Data availability statement.

Supplemental materials including computer code and output files for all simulations reported in this article are available at https://osf.io/4hgyt/

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65*, 241–261.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple indicator CT-C(*M*–1) model. *Psychological Methods, 8*, 38–60.

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. *Psychological Methods, 13*, 230–253.

Geiser, C. (2009). *Multitrait-multimethod-multioccasion modeling*. Munich, Germany: Akademische Verlagsgemeinschaft.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(*M*–1) model: A comment on Maydeu-Olivares & Coffman (2006). *Psychological Methods*, *13*, 49–57.

Geiser, C., Hintz, F. A., Burns, G. L., & Servera, M. (2019). Structural equation modeling of multiple-indicator multimethod-multioccasion data: A primer. *Personality and Individual Differences, 136*, 79-89.

Geiser, C., Koch, T., & Eid, M. (2014). Data-generating mechanisms versus constructively-defined latent variables in multitrait-multimethod analysis: A comment on Castro-Schilo, Widaman, and Grimm (2013). *Structural Equation Modeling, 21*, 509-523.

Geiser, C., & Simmons, T. G. (in press). Do method effects generalize across traits (and what if they don't)? *Journal of Personality*.

Grayson, D., & Marsh, H. W. (1994). Identification with deficient rank loading matrices in confirmatory factor analysis: Multitrait–multimethod models. *Psychometrika, 59*, 121–134.

Greiff, S., Fischer, A., Wüstenberg, S., Sonnleitner, P., Brunner, M., & Martin, R. (2013), A multitrait-multimethod study of assessment instruments for complex problem solving. *Intelligence, 41*, 579-596.

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling, 14*(3), 361-390.

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.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*(2), 109-133.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12*, 247-252.

Kenny, D. A., & Kashy, D. A. (1992). The analysis of the multitrait–multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*, 333–351.

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods, 7*(2), 228-244.

Marsh, H. W. (1989). Confirmatory factor analysis of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement, 13*, 335–361.

Marsh, H. W. (1993). Multitrait-multimethod analyses: Inferring each trait/method combination with multiple indicators. *Applied Measurement in Education, 6*, 49–81.

Marsh, H. W., & Grayson, D. A. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 177–198). Thousands Oaks, CA: Sage.

Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: Application of second-order confirmatory factor analysis. *Journal of Applied Psychology, 73*, 107–117.

Mazzetti, G., Schaufeli, W. B., & Guglielmi, D. (2018). Are workaholism and work engagement in the eye of the beholder? A multirater perspective on different forms of working hard. *European Journal of Psychological Assessment, 34*, 30-40.

McNeish, D., & Wolf, M. G. (2020). *Dynamic fit index cutoffs for confirmatory factor analysis models.* Manuscript submitted for publication. Retrieved October 1, 2020 from: https://psyarxiv.com/v8yru/download?format=pdf

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling, 19*, 86–98.

Muthén, L. K. and Muthén, B. O. (1998–2017). *Mplus User's Guide. Eighth Edition*. Los Angeles, CA: Muthén & Muthén.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analysing multitrait-multimethod data with structural equation models for ordinal variables applying the WLSMV estimator: What sample size is needed for valid results? *British Journal of Mathematical and Statistical Psychology. 59,* 195-213.

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research, 45*, 1-28.

Pohl, S., Steyer, R., & Kraus, K. (2008). Modelling method effects as individual causal effects. *Journal of the Royal Statistical Society. Series A, 171*, 41–63.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling, 25*, 21-40.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits: revised. *Annual Review of Clinical Psychology, 11*, 71-98.

van der Ende, J., Verhulst, F. C., & Tiemeier, H. (2020). Multitrait-multimethod analyses of change of internalizing and externalizing problems in adolescence: Predicting internalizing and externalizing DSM disorders in adulthood. *Journal of Abnormal Psychology, 129*, 343-354.

Widaman, K.-F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement, 9*, 1–26.

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80*, 379-405.