Degrees of Freedom in Multi-Group Confirmatory Factor Analysis: Are Models of Measurement Invariance Testing Correctly Specified?

Ulrich Schroeders

Psychological Assessment, University of Kassel

Timo Gnambs

Leibniz Institute for Educational Trajectories &

Johannes Kepler University Linz

Author Note

Ulrich Schroeders, Department of Psychology, University of Kassel, Germany; Timo Gnambs, Leibniz Institute for Educational Trajectories, Bamberg, Germany, and Johannes Kepler University Linz, Austria.

Correspondence concerning this article should be addressed to Ulrich Schroeders, Psychological Assessment, Institute of Psychology, University of Kassel, Holländische Str. 36-38, 34127 Kassel, Germany, E-mail: schroeders@psychologie.uni-kassel.de

Abstract

Measurement equivalence is a key concept in psychological assessment and a fundamental prerequisite for meaningful comparisons across groups. In the prevalent approach, multi-group confirmatory factor analysis (MGCFA), specific measurement parameters are constrained to equality across groups. The degrees of freedom (*df*) for these models readily follow from the hypothesized measurement model and the invariance constraints. In light of the current methodological crisis that questions the soundness of statistical reporting in psychology, we explored reporting inconsistencies in MGCFA invariance testing. We reviewed 128 studies from six leading peer-reviewed journals focusing on psychological assessment and recalculated the *df* for 302 measurement invariance testing procedures based on the information given in the publications. Overall, about a quarter of all articles included at least one reporting inconsistency with metric and scalar invariance being more frequently affected. We discuss moderators of reporting inconsistencies and identify typical pitfalls in invariance testing. Moreover, we provide example syntax for different methods of scaling latent variables and introduce a ShinyApp that allows for the recalculation of *df* in common MGCFA models to improve the statistical soundness of invariance testing in psychological research.

Keywords: reporting standards, reporting inconsistency, measurement invariance, structural equation modeling, degrees of freedom

Degrees of Freedom in Multi-Group Confirmatory Factor Analysis:
Are Models of Measurement Invariance Testing Correctly Specified?

Failures to replicate seemingly robust effects (Hagger et al., 2016; Simmons & Simonsohn, 2017; Wagenmakers et al., 2016) alongside questionable research practices that are commonly adopted in applied research (Simmons, Nelson, & Simonsohn, 2011) has put science in a state of turmoil, with psychology at its center. Fortunately, this methodological crisis has been understood not only as a threat but also as an opportunity to strengthen scientific conduct. In recent years, psychology as a discipline has begun to adopt a number of strategies to improve the robustness and trustworthiness of its findings (Chambers, 2017; Eich, 2014). These countermeasures include, among others, emphasizing statistical power (Bakker, van Dijk, & Wicherts, 2012), acknowledging uncertainty in statistical results (Cumming, 2014), undisclosing flexibility in data collection and analysis (Simmons et al., 2011), and distinguishing between exploratory and confirmatory data analysis (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Moreover, by adopting open practices such as making all material pertaining to a study including its questionnaires, experimental manipulations, raw data, and analyses scripts available to others, the replicability of the published findings are expected to increase (Nosek et al., 2015; Simonsohn, 2013). This transparency can be especially helpful to clarify why many peer-reviewed articles in psychology contain inconsistent statistical results that might impact the interpretation of its reported findings (Bakker & Wicherts, 2011; Cortina, Green, Keeler, & Vandenberg, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016). Recent reviews highlighted major weaknesses in the reporting of null-hypothesis significance tests (NHST) and structural equation models (SEM) that seriously undermine the trustworthiness of psychological science. In the present study, we review potential deficits in the statistical reporting of multi-group measurement invariance testing.

Reporting Inconsistencies in Statistical Results

Statistical results of journal articles are typically vetted by multiple peer reviewers and sometimes additionally statistical editors. Despite the thorough review process, many published articles contain statistical ambiguities. For example, Bakker and Wicherts (2011) scrutinized 281 articles from six randomly selected psychological journals (three with high and three with low-impact factor) and found around 18% of the statistical results incorrectly reported. Most recently, Njuiten and her colleagues (2016) revigorated this line of research by introducing the *R* package *statcheck* that automatically scans publications for reporting errors, that is, inconsistencies between a reported test statistic (e.g., *t*-value, *F*-value), the degrees of freedom (*df*), and its corresponding *p*-value. The sobering result of scanning over 250,000 publications of eight top-tier peer-reviewed journals (Nuijten et al., 2016) was that half of the articles contained at least one inconsistent *p*-value. Moreover, around 12% of the articles contained an inconsistency which changed the results significantly, often in line with the researchers' expectations. Even though the text recognition and evaluation routine has been criticized for being too sensitive (Schmidt, 2016), the study points to serious issues in the way researchers report their findings.

Considering the comprehensive methodological toolbox of psychologists, test statistics regularly used in NHST are comparatively simple. In applied research, more often sophisticated latent variable techniques are used to test structural hypotheses between several variables of interest. Recently, Cortina and colleagues (2017) reviewed 784 SEMs published in two leading organizational journals to examine whether the reported *df* matched the information given in the text. In case all necessary information was available to recalculate the *df* they only matched in 62% of the time. Reporting inconsistencies were particularly prevalent in structural (rather than measurement) models and were often large in magnitude. Thus, the trustworthiness of model evaluations seems questionable for a significant number of SEMs reported in the literature. In test and questionnaire development, methods used to examine the internal structure, to determine the reliability, and estimate the validity of measures typically also rely on latent variable modeling. The implementation of such procedures in standard statistical software packages also extends the spectrum of test construction—besides the traditional topics of reliability and validity—to other pressing issues such as test fairness and comparability of test scores across groups.

Measurement Invariance in Multi-Group Confirmatory Factor Analysis

Measurement invariance (MI) between two or more groups is given if individual differences in psychological tests can be entirely attributed to differences in the construct in question rather than membership to a certain group (see AERA, APA, & NCME, 2014). Thus, MI is an essential prerequisite to ensure valid and fair comparisons across cultures, administration modes, language versions, or sociodemographic groups (Borsboom, 2006b). Contemporary psychometric approaches to test for MI include various latent variable modeling techniques (e.g., Raju, Laffitte, & Byrne, 2002). In practice, multi-group confirmatory factor analysis (MGCFA) has become the *de facto* standard for testing MI in the psychological assessment literature, particularly for self-report instruments (see Putnick & Bornstein, 2016). Although different sequences can be implemented to test for MI (Cheung & Rensvold, 2002; Wicherts & Dolan, 2010), often a straightforward procedure of four hierarchical nested steps is followed (Millsap, 2011). In case constraining certain types of measurement parameters to equality leads to a considerable deterioration in model fit, the invariance assumption is violated. In the first step, *configural* MI, all model parameters except for necessary identification constraints are freely estimated across groups. For *metric* or *weak* MI, the factor loadings are constrained to invariance across groups allowing for comparisons of bivariate relations (i.e., correlations and regressions). In the third step, *scalar* or *strong* MI, the intercepts are set to be invariant in addition to the factor loadings. If scalar invariance holds, it is possible to compare the factor means across groups. In the last step, *strict* MI, additionally the item residuals are constrained to be equal across groups.

Depending on the chosen identification scheme for the latent factors (i.e., marker variable method, reference group method, and effects-coding method), different additional constraints have to be introduced (see Table 1): If a marker variable is selected for each latent factor, then its factor loading is fixed to 1, and its intercept is fixed to 0 in all MI steps outlined above. Alternatively, a reference group can be selected, which is sometimes preferred if the marker variable method exhibits convergence problems or choosing a marker variable will affect the results (Millsap, 2001). In practice, researchers frequently adopt a hybrid approach by fixing the factor loading of a marker variable to 1 and the mean of the latent variables in a reference group to 0 because this allows to interpret differences in factor means directly. Other identification schemes are possible and equally valid, but require different sets for identifying constraints. For example, Little, Slegers, and Card (2006) proposed a non-arbitrary way of identifying the mean and covariance structure by constraining the mean of the loadings to 1 and the sum of the intercepts to 0 for each factor. Importantly, the choice of identification constraints does not affect the number of estimated parameters or the results of the MI tests. To facilitate the implementation of MI testing, we provide example syntax for all the typical MI steps for all three methods of identification in *lavaan* (Rosseel, 2012) and *Mplus* (Muthén & Muthén, 1998-2017) in the supplemental material.

Table 1.

Constraints in MGCFA Tests for Measurement Invariance.

			Identification by Marker Variable			
		λ/ λ_m	τ/ τ_m	ε	$E(\xi)$	$Var(\xi)$
(1)	Configural invariance	*/ 1	*/ 0	*	*	*
(2)	Metric invariance	c/ 1	*/ 0	*	*	*
(3)	Scalar invariance	c/ 1	c/ 0	*	*	*
(4)	Strict invariance	c/ 1	c/ 0	c	*	*
			Identification by Reference Group			
		λ	τ	ε	$E(\xi)/ E(\xi^{(r)})$	$Var(\xi)/ Var(\xi^{(r)})$
(1)	Configural invariance	*	*	*	0/ 0	1/ 1
(2)	Metric invariance	c	*	*	0/ 0	*/ 1
(3)	Scalar invariance	c	c	*	*/ 0	*/ 1
(4)	Strict invariance	c	c	c	*/ 0	*/ 1
			Identification by Hybrid Approach			
		λ/ λ_m	τ	ε	$E(\xi)/ E(\xi^{(r)})$	$Var(\xi)$
(1)	Configural invariance	*/ 1	*	*	0/ 0	*
(2)	Metric invariance	c/ 1	*	*	0/ 0	*
(3)	Scalar invariance	c/ 1	c	*	*/ 0	*
(4)	Strict invariance	c/ 1	c	c	*/ 0	*

Note. λ = factor loading, λ_m = factor loading for marker variable, τ = intercept, τ_m = intercept for marker variable, ε = residual variance, $E(\xi)$ = latent factor mean, $E(\xi^{(r)})$ = latent factor mean in reference group, $Var(\xi)$ = latent factor variance, $Var(\xi^{(r)})$ = latent factor variance in reference group, * = parameter is freely estimated in all groups, c = parameter is constrained to equity across groups, 0/ 1 = parameter is fixed to a value of 0 or 1.

The Present Study

Given several critical reviews highlighting reporting inconsistencies in NHST and SEM (Bakker & Wicherts, 2011; Cortina et al., 2017; Nuijten et al., 2016), we were pursuing two objectives: First, we examined the extent of reporting inconsistencies in MI testing with MGCFA. Because the number of *df* for each MI step is mathematically determined through the hypothesized measurement model, we recalculated the *df* for the aforementioned MI steps based on the information provided in articles that were published in major peer-reviewed journals focusing on psychological assessment in the last 20 years. Second, we tried to identify potential causes for the misreporting (e.g., the complexity of the model or the used software packages). Furthermore, we highlight potential pitfalls when specifying the different steps of MI testing. To this end, we also provide example syntax for MI testing and introduce an easy to handle ShinyApp that allows double-checking the *df* in MI testing. Thus, the overarching aim is to improve the statistical soundness of MI testing in psychological research.

Method

Inconsistent *df* in MI tests of MGCFA were identified among issues of six leading peer-reviewed journals from the last 20 years (1996 to 2016) that regularly report on test development and questionnaire construction: *Assessment* (ASMNT), *European Journal of Personality Assessment* (EJPA), *Journal of Cross-Cultural Psychology* (JCCP), *Journal of Personality Assessment* (JPA), *Psychological Assessment* (PA), and *Personality and Individual Differences* (PAID). Studies were limited to reports of MGCFA that included one or more of the four MI steps outlined above. Not considered were single group tests of MI (i.e., longitudinal MI or multi-trait multi-method MI), second-order models, exploratory structural equation models, or MI testing with categorical data.

We first recalculated the *df* for all MI models from the information given in the text, tables, and figures (e.g., regarding the number of indicators, latent factors, cross-loadings). A configural model was coded as incorrect if the reported and recalculated *df* did not match. Then, the *df* for the metric, scalar, and strict MI model were also recalculated (syntax in the supplemental material) and compared to the reported *df*. In case, inconsistent *df* were identified at a specific step, the *df* for subsequent models were also recalculated by taking the reported (inconsistent) *df* of the previous step into account, which adopts a more liberal perspective. For example, if an author claimed to have tested metric invariance while also constraining the factor variances across all groups, this step was coded as incorrect. However, if in scalar MI testing the intercepts were additionally set to be invariant, this was coded as correct (despite the constrained factor variances). The coding was limited to the four types of MI as outlined above and we did not code partial MI. Both authors coded about half of the studies. In case inconsistent *df* were identified, the other author independently coded the respective study again. Diverging evaluations were discussed until a consensus was reached. We provide our coding sheets and all syntax within the *Open Science Framework* (Center for Open Science, 2017) at https://osf.io/6nh9d/?view_only=9228190ad66746ed9d3ade4bc8dd0b51

Results

We identified a total of 302 MI testing sequences that were published in 128 different research articles. Most articles were published in PA (31.3%) and PAID (23.4%), followed by EJPA (16.4%) and ASMNT (13.3%), whereas fewer articles were retrieved from JCCP and PA (7.8% each). The number of articles reporting MI testing within a MGCFA framework recorded a sharp increase in recent years. Nearly two-thirds of the articles were published within five years between 2012 and 2016 and over 88% within the last ten years (see Figure 1).

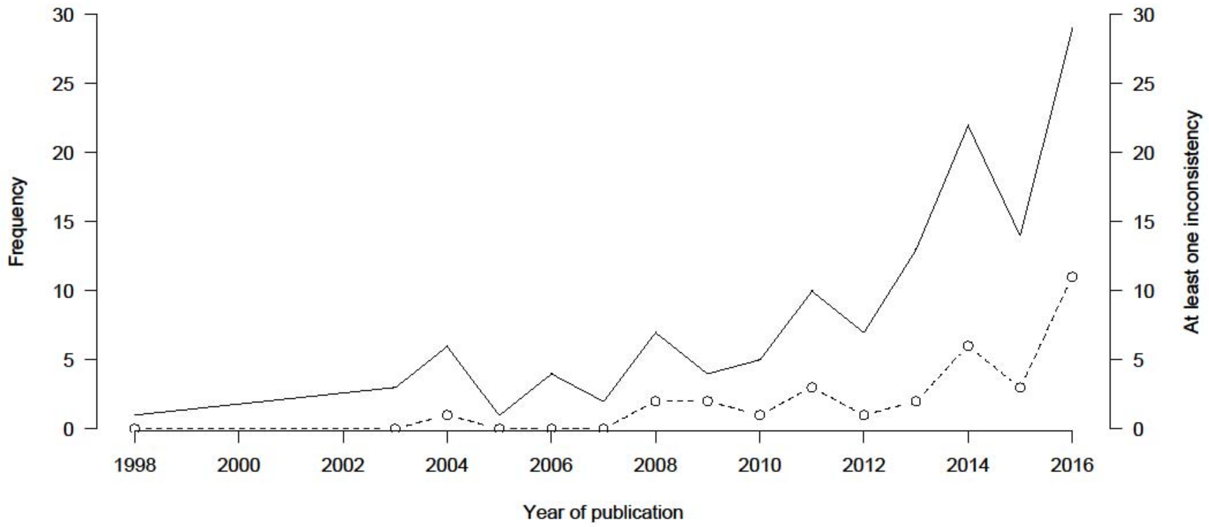

Figure 1. Studies Reporting Measurement Invariance Tests Over Time. *Note*. The solid line represents the number of studies reporting MI tests; dashed line represents the number of studies with at least one reporting inconsistency.

Out of 128 articles, 49 (38.3%) used *Mplus* to conduct MI testing, 24 (18.8%) used *LISREL*, and 23 (18.0%) used *Amos*. The remaining articles relied on specialized software such as *EQS* (n = 10) or *R* (n = 4), did not report their software choice (n = 17), or used more than one program (n = 1). On average, each article reported on 2.36 MI testing sequences (*SD* = 2.29, *Max* = 15). Further descriptive information on the model specification grouped by journal and publication year is summarized in Table S1 of the supplemental material.

Inconsistencies in Reported Degrees of Freedom

Half of the studies (48.4%) reported multiple MI tests (e.g., for age and sex groups); that is, the identified MI tests were not independent. Since variation was found on the study level rather than the MI test level (intra-class correlation = .995), we analyzed reporting inconsistencies on the level of studies rather than single tests of MI. Therefore, we aggregated the results to the article level and examined for each article whether at least one inconsistent *df* was identified for the different models in each MI step. The analyses revealed that out of 120 studies reporting configural MI, only 7 studies showed inconsistencies (5.8%, see Table 2). In contrast, tests for metric and scalar MI exhibited larger discrepancies between the reported and recalculated *df* (15.9% and 21.1%, respectively). Only one study reported incorrect *df* in strict MI.

Table 2.

Inconsistencies in Reported Degrees of Freedom.

	Configural	Metric	Scalar	Strict
Number reported [a]	120 (93.8)	126 (98.4)	95 (74.2)	40 (31.3)
Number inconsistent [b]	7 (5.8)	20 (15.9)	20 (21.1)	2 (0.1)

Note. [a] Percentages refer to 128 studies. [b] Percentages refer to the number of reported studies.

Table 3.

Predicting Occurrence of Inconsistencies Based on Study Characteristics

Predictors	*B*	*SE*	*z*	*OR*	95% CI	AME
(Intercept)	-3.66 *	0.80	-4.57	0.03	[0.00, 0.10]	
(1) Complexity of model	< 0.01 +	< 0.01	1.76	1.00	[1.00, 1.00]	.00 +
(2) Year of publication	0.22 *	0.08	2.68	1.25	[1.07, 1.49]	.03 *
(3) Journal (ref.: *Psychological Assessment, n* = 40)						
Assessment (*n* = 21)	1.07	0.85	1.25	2.91	[0.54, 16.14]	.13
European Journal of Psy. Assessment (*n* = 21)	0.21	0.85	0.24	1.23	[0.21, 6.53]	.02
Journal of Cross-Cultural Psychology (*n* = 10)	1.69 +	1.00	1.68	5.40	[0.72, 40.65]	.23
Journal of Personality Assessment (*n* = 10)	3.11 *	1.03	3.01	22.41	[3.19, 196.88]	.48 *
Personality and Individual Differences (*n* = 30)	1.38 +	0.72	1.91	3.97	[1.01, 17.77]	.17 +
(4) Software (ref: *Mplus, n* = 49)						
AMOS (*n* = 23)	1.85 *	0.79	2.33	6.33	[1.43, 34.72]	.23 *
EQS (*n* = 10)	3.70 *	0.99	3.73	20.54	[6.64, 347.88]	.57 *
LISREL (*n* = 24)	2.07 *	0.83	2.51	7.93	[1.68, 46.27]	.26 *
Remaining (*n* = 22)	1.42 +	0.86	1.66	4.14	[0.78, 24.78]	.16

Note. * $p < .05$; + $p < .10$. $n = 128$ studies. Logistic regression analysis with at least one inconsistency found (1) on study level versus not found (0) as an outcome. Predictors (1) and (2) were centered prior to analysis; predictors (3) and (4) were entered as dummy-coded variables. Nagelkerke's $R^2 = .37$. AME = Average marginal effects (Williams, 2012)

To shed further light on potential predictors of reporting inconsistencies, we conducted a logistic regression analysis using reporting inconsistency as an outcome (0 = no inconsistencies in *df*, 1 = at least one inconsistency in *df*). We added the (1) complexity of the model, (2) publication year, (3) journal, and (4) software package as predictors. Table 3 summarizes the respective results. The complexity of the model did not predict the occurrence of reporting errors. In contrast, the year of publication influenced the error rate with more recent publications exhibiting slightly more reporting inconsistencies. Given that most of the studies have been reported in recent years, the average marginal effect (AME; Williams, 2012) for an article including a reporting inconsistency was about 3.0% (*p* = .003) per year. Across all journals, a quarter of all published articles on MI included at least one *df* that we were unable to replicate (see dashed line in Figure S1 of the supplemental material). A comparison of the journals demonstrates subtle differences: In comparison to PA, the outlet that published most MI tests, JCCP (AME = 22.5%, *p* = .13) and PAID (AME = 17.4%, *p* = .05) reported slightly more inconsistent *df*. The highest rate of reporting inconsistencies between reported and recalculated *df* was found for JPA (AME = 48.3%, *p* = .001)—five of ten studies had inconsistencies. The most important predictor in the logistic analysis was the software package used in MI testing. In comparison to *Mplus*, studies using other software packages were more likely to report inconsistencies, that is, AMOS (AME = 22.3%, *p* = .02), LISREL (AME = 26.2%, *p* = .01), and most severely EQS (AME = 57.0%, *p* < .001).

Pitfalls in Testing Measurement Invariance

Without inspecting the analysis syntax of the reported studies, we can only speculate about the reasons for the identified reporting inconsistencies. However, in our attempts to replicate the *df* we spotted two likely sources of model specification: In testing metric MI, inconsistencies seem to have resulted in many cases (13 out of 20 flagged publications) from a misspecified model using the reference group approach for factor identification. As a reminder, the configural model includes fixing the variances of the latent variables to 1 in all groups, while freely estimating all factor loadings. The metric model, however, requires equality constraints on the factor loadings across groups, while relaxing constraints on the variances of the latent variables except for the reference group. It seems that many authors neglected to free the factor variances and, thus, instead of testing a metric MI model, evaluated a model with invariant loadings and variances.

Issues in reporting scalar MI can in many instances (12 out of 20 flagged studies) be traced back to a misspecified mean structure. SEM is a variance-covariance based modeling approach, and in a single group case, researchers are usually not interested in the mean structure. Therefore, scalar MI tests, in which the mean structure plays a vital role, seems to present particular difficulties and make up for the largest number of reporting inconsistencies. Again, we suspect that researchers adopting the reference group or hybrid approach for factor identification neglected to free previously constrained latent factor means (see Table 1). As a result, instead of testing for scalar MI, these models in fact evaluated invariant intercepts and means fixed to 0 across groups. Such model misspecifications are not trivial and have severe consequences for model fit evaluations: In a simulated MGCFA MI example, we compared a correctly specified scalar MI model with freely estimated latent factor means (except for the necessary identifying constraint) to a model, in which all factor means were fixed to zero.

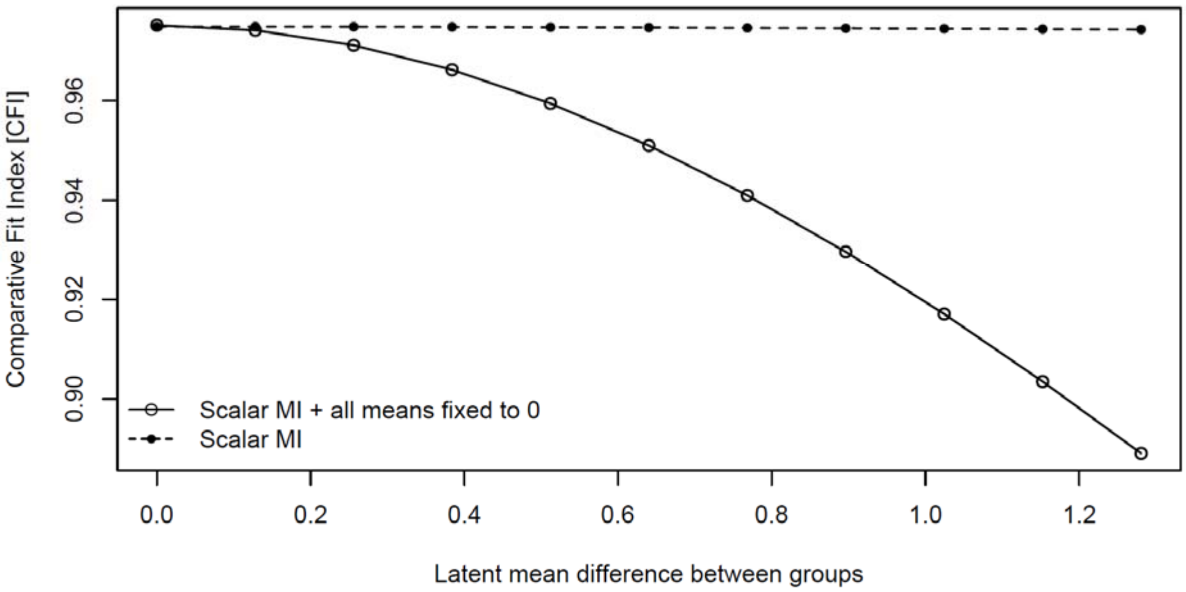

Figure 2. Consequences of Fixing the Means to 0 in Scalar Measurement Invariance Testing on Model fit.

Figure 2 demonstrates that already moderate differences in the latent means (*d* ≈ .50), result in a drop in the comparative fit index (CFI) from an initially good fitting model (CFI = .98) to values below what is usually considered acceptable (CFI ≥ .95). Thus, if the means are constrained to zero, any differences in the latent means are passed on to the intercepts; if these are also constrained to equality, the unmodeled mean differences can result in a substantial model deterioration. As a consequence, misspecified scalar MI models can lead to serious misinterpretation, that is, the rejection of the scalar MI model.

Discussion

The concept of measurement equivalence is pivotal for psychological research and practice. To address substantial research questions, researchers depend on information about the psychometric functioning of their instruments across sex and ethnic groups, clinical populations, etc. Accordingly, reporting issues in MI testing are not restricted to a specific field but affect different disciplines such as clinical and I/O psychology. The extent of inconsistencies found in the psychological assessment literature was rather surprising: One out of four studies reporting MI tests included an incorrectly specified or, at least, insufficiently described model. Thus, a substantial body of literature on the measurement equivalence of psychological instruments seems to be questionable at best or inaccurate. This percentage is probably a lower boundary of the true error rate due to the way we coded the MI tests (i.e., no subsequent errors, exclusion of studies that specified different configural models across groups). Since our analysis was limited to inconsistencies in the *df*, it is possible that additional errors may have occurred (e.g., handling of missing data, incorporating nested structures, or using different estimators that might be more appropriate for categorical data). To identify these and similar flaws, both the raw data and the analyses scripts would be necessary to reanalyze the data.

Regarding the cause of reporting inconsistencies, the results of the logistic regression provide us with some valuable clues: The increased popularity of MGCFA MI testing in psychological research was accompanied by an increase in reporting inconsistencies. This is not an unusual pattern in the dissemination of psychological methods: After the formal (and often formalized) introduction of a new method by psychometricians more and more users adopt and apply the method—sometimes without a deeper understanding of the underlying statistics. However, the strongest effect on reporting issues was observed for the software package used to conduct MI tests. In comparison to *Mplus*, other software packages performed worse, which might be due to the extensive documentation and training materials. Or, it can more likely attributed to a selection effect, because more advanced users prefer scripting languages. Taken together, we think that the results of the logistic regression may point to a general problem with the formal methodological and statistical training of psychologists (Borsboom, 2006a).

Recommendations for Reporting MGCFA MI Testing

In the following, some recommendations are given to improve the accuracy of reporting statistical results in the framework of MI reporting. These recommendations apply to all parties involved in the publication process—authors, reviewers, editors, and publishers:

First, make sure that all necessary information concerning the measurement model is described. This pertains not only to the specification of the number of indicators, factors, cross-loadings, residual covariances, and groups but also to the constraints introduced at the different MI steps. It should be explicitly stated which parameters were constrained and which constraints were relaxed (e.g., in the notes of a table), so that it is clear which models are nested within each other. In addition, model fit indices (including *df*) for all invariance steps should be reported.

Second, use unambiguous terminology when referring to specific steps in MI testing. In our reading of the literature, we found several cases, in which the description in the method section did not match the restrictions given in the respective table. One way to clarify which model constraints have been introduced is to label the invariance step by the parameters that have been fixed (e.g., "invariance of factor loadings" instead of "metric invariance").

Third, in line with the recommendations of the *Association of Psychological Science* (Eich, 2014) and the extensive efforts of the *Open Science Framework* (Nosek et al., 2015) to make scientific research more transparent, open, and reproducible, we strongly advocate to make the raw data and the used analysis syntax available in freely accessible data repositories. As a pleasant side-effect, there is also evidence that sharing detailed research data is associated with increased citation rate (Piwowar, Day, & Fridsma, 2007). If legal restrictions or ethical considerations prevent the sharing of raw data, it is possible to create synthesized data sets (Nowok, Raab, & Dibben, 2016).

Fourth, we encourage authors and reviewers to routinely double-check the *df* of the reported models. In this context, we welcome the recent effort of journals in psychology to include soundness checks on manuscript submission by default to improve the accuracy of statistical reporting. To this end, one may refer to the supplemental material that includes example syntax for all steps of MI in *lavaan* and *Mplus* for different ways of scaling latent variables or use our ShinyApp to double-check the *df* of the different MI steps for a given model (https://psychresearch.shinyapps.io/df_in_mi/).

Fifth, statistical and methodological courses need to be taught more rigorously in university teaching, especially in structured Ph.D. programs. A vigorous training should include both conceptual (e.g., Borsboom, 2006b; Markus & Borsboom, 2013) and statistical work (e.g., Millsap, 2011). To bridge the gap between psychometric researchers and applied working psychologists, a variety of teaching resources can be recommended that introduce invariance testing in general (Cheung & Rensvold, 2002; Wicherts & Dolan, 2010) or specific aspects of MI such as longitudinal MI (Geiser, 2013), and MI with categorical data (Pendergast, von der Embse, Kilgus, & Eklund, 2017).

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*, 666–678. doi:10.3758/s13428-011-0089-5

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, *71*, 425–440. doi:10.1007/s11336-006-1447-6

Borsboom, D. (2006b). When does measurement invariance matter? *Medical Care*, *44*, 176–181. doi:10.1097/01.mlr.0000245143.08679.cc

Chambers, C. (2017). *The seven deadly sins of Psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. doi:10.1207/S15328007SEM0902_5

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test? *Organizational Research Methods*, *20*, 350–378. doi:10.1177/1094428116676345

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*, 3–6. doi:10.1177/0956797613512465

Geiser, C. (2013). *Data Analysis with Mplus*. New York: Guilford Press.

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Bruyneel, S. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573. doi:10.1177/1745691616652873

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*, 59–72. doi:10.1207/s15328007sem1301_3

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York: Routledge.

Millsap, R. E. (2001). When trivial constraints are not trivial: the choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 1–17. doi:10.1207/S15328007SEM0801_1

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. doi:10.1126/science.aab2374

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*. doi:10.18637/jss.v074.i11

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. doi:10.3758/s13428-015-0664-2

Pendergast, L., von der Embse, N., Kilgus, S., & Eklund, K. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology*, *60*, 65-82. doi:10.1016/j.jsp.2016.11.002

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLOS ONE*, *2*, e308. doi:10.1371/journal.pone.0000308

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. doi:10.1016/j.dr.2016.06.004

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529. doi:10.1037/0021-9010.87.3.517

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. doi:10.18637/jss.v048.i02

Schmidt, T. (2016). Sources of false positives and false negatives in the STATCHECK algorithm: Reply to Nuijten et al.(2016). *ArXiv Preprint ArXiv:1610.01010*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., & Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, *28*, 687–693.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*, *24*, 1875–1888. doi:10.1177/0956797613480366

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., … Blouin-Hudon, E.-M. (2016). Registered Replication Report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638. doi:10.1177/1745691612463078

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice*, *29*, 39–47. doi:10.1111/j.1745-3992.2010.00182.x

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal*, *12*, 308–331.