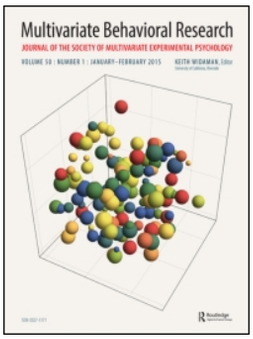

Routledge
Taylor & Francis Group

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

Multilevel Factor Analysis: Reporting Guidelines and a Review of Reporting Practices

Eun Sook Kim, Robert F. Dedrick, Chunhua Cao & John M. Ferron

To cite this article: Eun Sook Kim, Robert F. Dedrick, Chunhua Cao & John M. Ferron (2016) Multilevel Factor Analysis: Reporting Guidelines and a Review of Reporting Practices, Multivariate Behavioral Research, 51:6, 881-898, DOI: 10.1080/00273171.2016.1228042

To link to this article: https://doi.org/10.1080/00273171.2016.1228042

MULTIVARIATE BEHAVIORAL RESEARCH
2016, VOL. 51, NO. 6, 881–898
http://dx.doi.org/10.1080/00273171.2016.1228042

Routledge
Taylor & Francis Group

QUANTITATIVE METHODS IN PRACTICE: TUTORIAL

Multilevel Factor Analysis: Reporting Guidelines and a Review of Reporting Practices

Eun Sook Kim, Robert F. Dedrick, Chunhua Cao, and John M. Ferron

University of South Florida

ABSTRACT

We provide reporting guidelines for multilevel factor analysis (MFA) and use these guidelines to systematically review 72 MFA applications in journals across a range of disciplines (e.g., education, health/nursing, management, and psychology) published between 1994 and 2014. Results are organized in terms of the (a) characteristics of the MFA application (e.g., construct measured), (b) purpose (e.g., measurement validation), (c) data source (e.g., number of cases at Level 1 and Level 2), (d) statistical approach (e.g., maximum likelihood), and (e) results reported (e.g., intraclass correlations for indicators and latent variables, standardized factor loadings, fit indices). Results from this review have implications for applied researchers interested in expanding their approaches to psychometric analyses and construct validation within a multilevel framework and for methodologists using Monte Carlo methods to explore technical and methodological issues grounded in realistic research design conditions.

KEYWORDS

Multilevel factor analysis; multilevel confirmatory factor analysis; multilevel exploratory factor analysis; reporting guidelines; systematic review

In recent years, there has been a rapid increase in the application of multilevel or hierarchical models across a wide range of disciplines. Multilevel models are characterized by a nested data structure in which there are multiple levels or units of analysis (e.g., students nested within classrooms nested within schools; repeated measures nested within individuals). Reviews of methodological issues and reporting practices of results from these models have been conducted in the fields of school psychology (Graves & Frohwerk, 2009), rehabilitation psychology (Jackson, 2010), education, and the social sciences (Dedrick et al., 2009; Schreiber & Griffin, 2004), but much of the focus of these reviews has been on what Sirotnik (1980) has called the *study phase* of the research process as opposed to the *psychometric phase*. Sirotnik defined the *study phase* as "those efforts which take the measures as given and pursue statistical analyses relevant to the research questions" whereas the *psychometric phase* is "aimed at assessing the qualities of the measurement devices" (p. 245). Sirotnik in an early review of 40 studies of organizational climate found that even when researchers either directly or indirectly addressed the unit of analysis issue in the study phase of the research process, most conducted psychometric analyses such as exploratory factor analysis and reliability analysis of individual-level data ignoring the organizational unit of analysis. Sirotnik concluded from these results that "the psychometric implications of unit-of-analysis issues have been almost universally ignored in the organizational climate" (Sirotnik, 1980, p. 258).

Since the initial work of Sirotnik (1980), there has been a growing effort to address the unit-of-analysis issue in the conceptualization and validation of educational and psychological measures and connect the psychometric and study phases of the research process. Early work in this area was conducted by Raudenbush, Rowan, and Kang (1991), who used a three-level hierarchical linear model (HLM) to analyze the psychometric properties of five dimensions of school climate (principal leadership, staff cooperation, teacher control, teacher efficacy, and teacher satisfaction). Raudenbush et al. conducted psychometric analyses of a 35-item Likert scale questionnaire (item responses represent Level 1) completed by 1,967 high school teachers (individuals represent Level 2) from 110 schools (Level 3). Three-level HLM analyses were used to partition the total variation of the scores into item-level, individual teacher-level, and school-level components. Raudenbush et al. (1991) then connected the psychometric phase of the analyses to the study phase by examining the relation between Level 2 teacher factors (e.g., teacher sex) and Level 3 school factors (e.g., school size) and school climate. A similar application using the three-level HLM framework was conducted by Miller and Murdock (2007) to measure the classroom climate

CONTACT Eun Sook Kim ekim3@usf.edu Department of Educational and Psychological Studies, University of South Florida, 4202 E. Fowler Avenue, EDU 105, Tampa, FL 33620.

Supplemental data for this article can be accessed on the publisher's website.

(mastery goal structure, performance goal structure, teacher competence, teacher respect, teacher interest, and teacher expectancy measured with 39 Likert items) perceived by 689 students from 57 high school classes.

Within the three-level HLM framework presented by Raudenbush et al. (1991), items are equally weighted and combined into their appropriate scales. By equally weighting all items on a scale and rescaling the item-level error variances to have equal variances, the Level 1 model in HLM can be conceptualized as a confirmatory factor analysis model with factor loadings constrained to be equal. While this model provides a means for conducting psychometric analyses within a multilevel framework, the assumptions underlying the model (i.e., parallel tests that assume the same true score and the same error variances for the measures) are strong and difficult to meet in applied research. An alternative to the three-level HLM of Raudenbush et al. (1991) was presented by Muthén (1991, 1994) in the form of a latent variable multilevel factor analysis model (see also Hox, 2002; Kamata, Bauer, & Miyazaki, 2008). This two-level factor model is part of a comprehensive and flexible framework that can be used to evaluate multilevel constructs and link the psychometric and study phases of the research process. The framework, as implemented via the software program M*plus* (Muthén & Muthén, 1998–2012), includes both exploratory and confirmatory factor models that can be used with equal or unequal (unbalanced) sample sizes of Level 1 units (e.g., individuals) across Level 2 units (e.g., organizations). Whereas factor analysis at a single level analyzes the total covariance matrix of the observed variables, multilevel factor analyses decompose the total sample covariance matrix into pooled within-group and between-group covariance matrices and uses these two matrices in the analyses of the factor structure at each level.[1] With multilevel factor analyses, it is possible to evaluate a variety of alternative or competing models including those that have (a) the same number of factors at each level and equal loadings across levels, (b) the same factor structures but different loadings across levels, or (c) a different number of factors at the two levels. Evaluation of alternative factor models using multilevel factor analysis provides a way of bringing together empirical evidence and theory in the construct validation process of multilevel constructs.

As presented in Table 1, there is a growing number of applications of multilevel factor analysis (MFA) that have been driven both by the heightened awareness of the unit-of-analysis issue in psychometric analyses and by the availability of statistical software such as M*plus* (Muthén & Muthén, 1998–2012), LISREL (Jöreskog & Sörbom, 2006), EQS (Bentler, 2000–2008), and MLwiN (Rasbash, Steele, Browne, & Goldstein, 2012). With the increased use of MFA in the construct validation process, it is an opportune time to pause and examine applied researchers' use of the relatively new technique of MFA, including their rationale of using the approach, the results reported from the analyses, and the problems encountered during the analyses. Limited information is available on these and other issues, and therefore one of the purposes of this study was to explore current uses and reporting practices of MFA via a systematic review of 72 published applications of multilevel factor analysis using reporting guidelines for MFA developed in this study. A description of current practices in the use of MFA in the research literature may sensitize methodologists to technical and methodological issues that could be explored through simulation research (for an example, see Hox & Maas's (2001) simulation study that examined four conditions: balanced versus unbalanced groups; number of groups; average group size; and level of the intraclass correlation). In addition, an exploration of what researchers are reporting related to their conceptualization of multilevel constructs and how they address the unit of analysis issue may provide guidance to other researchers attempting to gain insight into the meaning of their constructs and the connection to the construct validation process.

Table 1. Number of multilevel factor analysis articles coded by year and discipline ($N = 72$).

Year	*N* (%)	Discipline	*N* (%)
1994–2000	2 (2.8)	Education	24 (33.3)
2001–2005	12 (16.7)	Psychology	22 (30.5)
2006–2010	26 (36.1)	Health/Nursing	18 (25.0)
2011–2014	32 (44.4)	Exercise Science	4 (5.6)
		Sociology	3 (4.2)
		Management	1 (1.4)

Note. Some applications are related to both education and psychology.

Purposes

This study had three primary purposes: (a) provide reporting guidelines for MFA, (b) describe the current use and reporting practices of MFA in the published literature, and (c) present a summary of the reported fit statistics and parameter estimates from these multilevel factor analyses. Although guidelines for single-level exploratory and confirmatory factor analysis (Bandalos & Finney, 2010; DiStefano & Hess, 2005; Henson & Roberts, 2006; Jackson, Gillaspy, & Purc-Stephenson, 2009) have been presented along with reporting practices for structural equation modeling (McDonald & Ho, 2002) and multilevel modeling (Ferron et al., 2008; McCoach, 2010), there are no reporting guidelines specific to multilevel factor analysis. Thus, the first purpose of this study was to present guidelines for reporting on MFA.

[1] The current version of Mplus (7.4) has the THREELEVEL option and supports the raw data entry for data analysis as well as a covariance or correlation matrix.

The second purpose of this study was to determine to what extent there is a gap between these reporting guidelines and what was reported by the authors of the MFA applications. Reviews of reporting practices of various statistical approaches have frequently found substantial gaps between recommended reporting practices and what is presented by researchers (Dedrick et al., 2009; DiStefano & Hess, 2005; Keselman et al., 1998; Thoemmes & Kim, 2011). As part of this analysis, we examined the degree to which researchers were transparent in the reporting of their research. The American Educational Research Association's (AERA's) Standards for Reporting on Empirical Social Science Research in AERA Publications (American Educational Research Association, 2006) have stressed the importance of transparency in the reporting of research. This transparency is achieved by making explicit all aspects of the research process, including the (a) data collection design and the logic of the analyses in relation to the study's purpose(s), (b) assumptions underlying the analyses and the extent to which these assumptions are met, and (c) procedures used in the analyses (e.g., estimation method) and "any considerations during the data analysis that might compromise the validity of the statistical analyses" (p. 37). Transparency allows other researchers to critically examine and build on the research. An understanding of the degree of transparency with current applications of multilevel factor analyses is particularly important given the complexity of these models and their increasing application across disciplines. To this end, we (a) identified multilevel factor analysis applications in the published literature; (b) described the purposes of these applications (i.e., researchers' rationale for using the technique), including details related to the constructs being examined, data collection designs, types of data (e.g., if questionnaires, the number of items and response scale, and number of dimensions; (c) identified the statistical approaches to conducting MFA (e.g., Muthén's Multiple-Step process); and (d) described what results are presented from these analyses (e.g., intraclass correlations, fit statistics, factor loadings, and standard errors) and the format used to present these results (e.g., two-level schematic diagram).

The third purpose of the study was to provide summary statistics of MFA results presented in the reviewed applications. These MFA results included indices of data dependency such as intraclass correlations (ICCs) and design effects, model fit indices, and parameter estimates such as standardized factor loadings, factor correlations, and residual variances at each level. The collected information can be used not only for applied researchers to appraise their methodological approaches and outcomes in comparison with the extant literature but also for methodologists interested in designing simulation studies that reflect authentic data and real research settings.

Reporting guidelines for multilevel factor analysis

The proposed reporting guidelines for multilevel factor analysis are presented in Table 2. The extant guidelines for the reporting of factor analysis (e.g., Bandalos & Finney, 2010) and multilevel modeling (e.g., Ferron et al., 2008; McCoach, 2010) along with the AERA's Standards (American Educational Research Association, 2006) were used as the framework for developing the MFA reporting guidelines. Because many of these guidelines were explicated in the aforementioned articles, we do not reiterate those points here. Instead, several issues specific to MFA are discussed in detail in the following section, which is organized under six subsections: (a) multilevel constructs, (b) model representation via schematic diagram, (c) analytic procedure, (d) cross-level invariance, (e) level-specific reliability, and (f) model fit evaluation.

Multilevel constructs

As noted by Sirotnik (1980), "the unit-of-analysis problem is not a statistical problem" (p. 246). He continued by saying that the determination of the appropriate psychometric analysis "must be based on substantive consideration" that included whether what was being measured was conceptualized as "fundamentally systemic (i.e., intrinsic to the group) … [or] as fundamentally phenomenological (i.e., intrinsic to the individual)" (p. 246). More recently, Zumbo and Forer (2010) and Zumbo, Liu, Wu, Forer, and Shear (2010) have provided a framework for multilevel construct validation that focuses on the substantive meaning of the constructs, especially in situations in which data are collected at a lower level of analysis (e.g., students) but inferences are made about a higher-level unit of analysis (e.g., school).

Building on the framework of Zumbo et al. (2010), Stapleton, Yang, and Hancock (2016) have introduced four types of multilevel constructs on the basis of the level at which the constructs are conceptually meaningful. The four types of multilevel constructs include a (a) Level 1 construct with the individual as the unit of interest, (b) shared cluster construct, (c) configural cluster construct, and (d) simultaneous shared and configural cluster construct. The shared and configural cluster constructs are also called reflective (or composition) and formative (or compilation) constructs, respectively (Bliese, 2000; Lüdtke et al., 2008).

When individuals are the units of interest and the construct is conceptually at Level 1, Stapleton et al. (2016) have argued that no true construct may exist at the second level, and thus, specifying a factor structure at Level 2 may not be appropriate. To estimate this type of

Table 2. Reporting guidelines for multilevel factor analysis.

Purposes
1. Purposes/research questions of the study are presented.
2. Theory and literature review are pertinent to the study purposes.
3. Type of factor analysis (exploratory and confirmatory) is specified and justified.
i) Characteristics of multilevel factor analysis
4. Rationale of multilevel approach is discussed and justified.
5. Structure of multilevel data (e.g., individuals nested within organizational units or repeated measures nested within individuals) is identified.
6. Construct measured at each level is identified and theoretically supported.
7. Latent variable model at each level (e.g., number of factors, factor structure such as a higher-order model) is clearly specified preferably through equations and schematic diagrams in addition to detailed verbal descriptions.
8. In EFA, the methods of factor extraction (e.g., Geomin) and rotation (orthogonal or oblique) are presented and justified.
9. In EFA, the number of factors extracted at each level and the criteria used for the decision are presented and discussed.
10. The number and type of items (e.g., continuous or categorical; if categorical, the number of response categories) are reported along with the psychometric qualities of the items and measures. The full set or example items can be introduced either in text or in appendix.
ii) Data Source
11. The unit of analysis at each level is identified and the corresponding sample size and cluster size in addition to the total sample size are reported and justified (e.g., power analysis at each level).
12. The sampling method at each level is described. Sampling weights, if appropriate, are applied.
13. Missing data mechanism including complete data and missing data treatment at each level are discussed.
14. Any manipulation of data (e.g., transformation or item parceling) are discussed and justified.
iii) Statistical approach
15. The type of matrix analyzed is specified.
16. The software and version used are identified.
17. Statistical assumptions (e.g., multivariate normality if applicable) are checked and discussed.
18. The procedures of multilevel factor analysis such as Muthén's Five-Step Approach are clearly described.
19. The estimation method is discussed.
20. Estimation problems (e.g., nonconvergence, negative level-2 residual variance) and the method of handling the problems, if present, are explained.
21. If multiple competing models are introduced, model comparison methods (e.g., likelihood ratio test or information criteria) are presented and discussed.
22. Model modification, if relevant, is reported and justified. The criteria and methods of modification are presented and discussed.
iv) Results and interpretations
23. Model evaluation criteria and relevant statistics (e.g., fit indices such as SRMR-between and SRMR-within) are presented and discussed.
24. Relevant parameter estimates (e.g., pattern/structure coefficients in EFA, standardized/unstandardized factor loadings in CFA, factor correlations), standard errors, and measures of variance accounted for (e.g., *R*-square) are reported at each level and interpreted.
25. Level-specific scale reliability is reported and discussed.
26. The intraclass correlations (ICC) of items are reported and interpreted. If appropriate (i.e., cross-level invariance holds), factor ICCs are reported and interpreted.
27. If theoretically relevant, cross-level invariance is evaluated and discussed.
28. Factors are interpreted on the basis of theory.
29. If needed, factor scores are estimated at each level separately.
30. The limitations of the study are presented and discussed.

construct from multilevel data, Stapleton and colleagues recommend a design-based approach in which standard errors are adjusted to take into account data dependency or a multilevel CFA with a saturated model at Level 2. A shared cluster construct is fundamentally a Level 2 construct but is measured by individuals' item responses (e.g., teacher instructional quality perceived by students). In this case, individuals are informants of the construct and thus are interchangeable sources of information about the cluster construct. Individuals' responses within groups for a shared construct should exhibit strong within-group agreement. On the other hand, a configural cluster construct is an aggregate of individual characteristics such as school-average SES that is composed of student SES (Lüdtke et al., 2008). Individuals' responses within groups for a configural construct are not interchangeable, and thus responses may disagree within clusters. Finally, shared constructs could exist representing contextual effects in addition to configural constructs.

Different types of multilevel constructs require different model specifications. For example, if a Level 1 construct is of focal interest, a saturated model can be specified at the between level (Level 2) whereas for a configural cluster construct, it is suggested that factor loadings be constrained to be equal across levels (Stapleton et al., 2016). Accordingly, the interpretation of parameter estimates and related statistics such as scale reliability is more meaningful at a certain level than the other depending on the type of multilevel construct. Thus, analysts need to conduct construct validation with multilevel models that are consistent with the multilevel conceptualization of the construct. That is, before jumping into MFA, researchers need to decide at which level the construct of interest is meaningful, what specific type of construct they are working with, and subsequently how the model of the construct should be specified and estimated. These decisions should be explicitly articulated in the researchers' papers. Lack of information on the conceptualization of

the multilevel constructs in a study hinders readers from appraising the adequacy of model specification and interpretation.

Model representation via schematic diagram

Model specification is a central part of the research process that has some added complexities when working with models with multiple levels. Researchers commonly use one or more approaches to communicate their models, including using diagrams, equations, matrix representations, and verbal descriptions. Although equations and matrix representations are highly efficient in summarizing multilevel factor models, graphical representations can facilitate the communication of complex models, particularly for broad audiences that include readers from nontechnical backgrounds. Conventionally, in factor analysis (FA) diagrams, squares and circles represent observed and latent variables, respectively. A directional relation from one variable to another is denoted by an arrow. Variances and covariances are represented by curved, double-headed arrows. Due to model complexity (e.g., decomposition of variance into within and between components; random and fixed effects), multilevel FA diagrams entail more refined elements. However, there are variations in depicting the special elements of multilevel models that will be explained with examples below.

Figure 1 presents five different diagrams used to represent multilevel FA models in the literature. One strength of the format in Figure 1, part (a), is that it is possible to visualize the decomposition of the total variance of the observed variables into variance due to the latent variables at the between-group level, residual variance at the between-group level, variance due to the latent variables at the within-group level, and residual variance at the within-group level. In Figure 1, part (a), an observed variable (a square) is decomposed into two latent variables (two dotted circles associated with an observed variable)—a within-group estimate as a deviation score from the group mean at the within level and the estimated group mean (or intercept) at the between level. The decomposed estimate of an observed variable at each level is predicted by the latent factor (an arrow from a circle) with an error (or residual; a short or slanted arrow) at the corresponding level. Similar diagrams are illustrated by Stapleton et al. (2016). By contrast, in Figure 1, parts (b) and (c), this decomposition is not evident because

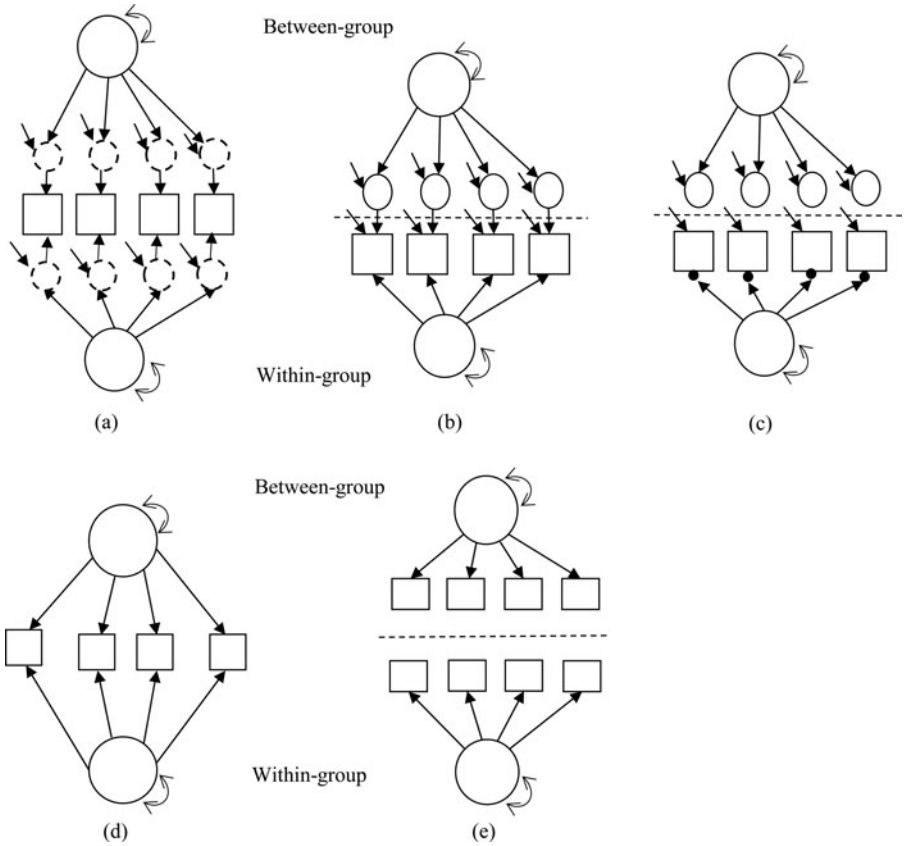

Figure 1. Path diagram of multilevel factor analysis. Square = observed variable; circle = unobserved variable; dotted circle = unobserved variable due to the decomposition of the corresponding observed variable into within-group and between-group components; filled dot = random intercept (or mean); curved arrow = variance; short slanted arrow = residual or residual variance.

the decomposed within-level component of an observed variable is missing. However, the arrows that connect the between-level items to the respective within-level items in Figure 1, part (b), and the filled dots at the end of the arrows in Figure 1, part (c), depict that the intercepts (or means) of items vary across clusters (groups). In Figure 1, parts (d) and (e), the element of random intercepts varying across clusters and the decomposition of observed variables are not communicated. As a side note, using filled dots for random effects is convenient to represent random slopes although random slopes are not common in multilevel FA. That is, filled dots *on* the arrows (i.e., slopes) represent random slopes.

It should be noted that some elements of the diagrams can be suppressed for simplicity when those elements are the default of a model (e.g., variance of an exogenous variable and residual variance). For example, residuals that are represented by circles at the end of short or slanted arrows are not shown in any diagram in Figure 1. Furthermore, it is not uncommon to hide even the slanted arrows (the whole unexplained components of observed variables) because errors are assumed to be present for endogenous variables. However, when these elements are specified differently from the default, the specification should be clearly represented in the diagram (e.g., between-level residual variance constrained at zero, residual covariance). In summary, parts (d) and (e) of Figure 1 are not appropriate for multilevel FA because they fail to describe the variance decomposition of observed variables. Figure 1, part (a), which clearly depicts the variance decomposition, is recommended to represent multilevel FA models, and when appropriate, the curved double-headed arrow (factor variance) and slanted arrows (residuals) can be suppressed for simplicity.

Analytic procedure

As an MFA model is multifaceted, the procedure for testing the model is also intricate and varies depending on what information about the multilevel constructs is sought by the researcher. One procedure suggested in the MFA literature is Muthén's (1994) five-step approach. Step 1 involves conducting a confirmatory factor analysis (CFA) on the sample total covariance matrix (traditional single-level CFA); Step 2 entails computation of intraclass correlations to explore between-group variability and the viability of multilevel analyses; Step 3 conducts a CFA on the pooled within-covariance matrix; Step 4 conducts the CFA on the sample between-group matrix; and Step 5 is the full multilevel analysis that involves a simultaneous analysis of the within- and between-group covariance matrices to evaluate the factor structure at each level. Although Muthén framed this approach under confirmatory factor analysis, a similar multistep approach is well suited to multilevel exploratory factor analysis (EFA) and is illustrated by Reise, Ventura, Nuechterlein, and Kim (2005) and van de Vijver and Poortinga (2002). That is, factor solutions are explored at each level separately with pooled within- and between-group matrices when ICCs are considered nontrivial. In this case, standard procedures of single-level EFA can be employed for the EFA at each level, and thus, it is suggested that researchers follow reporting guidelines developed for single-level EFA (e.g., Bandalos & Finney, 2010).

Cross-level invariance

Cross-level invariance (also called factor invariance across levels) is essential for the construct validity of the configural cluster construct because the construct at Level 2 "merely reflects the cluster aggregate of the individual construct at Level 1" (Stapleton et al., 2016). Cross-level invariance is also expected when a construct is at the individual level but there are some (spurious) clustering effects to be modeled at Level 2 because the Level 2 relations simply reflect the Level 1 relations (Stapleton et al., 2016; Zyphur, Kaplan, & Christian, 2008). In the measurement invariance literature, cross-level invariance is viewed as evidence of metric invariance across clusters (Jak, Oort, & Dolan, 2013). In other words, if metric invariance holds or Level 1 factor loadings are equal across clusters, then the factor loadings at Level 2 should be identical to those at Level 1. Thus, Jak et al. (2013) interpreted the violation of cross-level invariance as cluster bias (i.e., noninvariant factor loadings across clusters). In addition, the computation of the factor ICC requires the equality of factor loadings across levels because the metric used at Level 1 and Level 2 should be consistent to take a ratio of Level 2 factor variance to the total factor variance (i.e., sum of Level 1 and Level 2 factor variances).

For these reasons, it is important to report whether cross-level invariance is achieved and how the invariance is tested. It should be kept in mind that when factors are not invariant across levels, the factor ICC cannot be estimated; if needed, factor scores should be estimated at each level because factor scores cannot be pooled across levels due to different relations of factors with indicators across levels. Analogously, the meaning of factors specified in a CFA or extracted in an EFA should be discussed at each level.

Level-specific reliability

Given that reliability is defined as true score variance over total observed variance (McDonald, 1999), estimating

reliability with multilevel data is complex because variances are decomposed into within- and between-group components. Geldhof, Preacher, and Zyphur (2014) have advocated level-specific reliability—estimating reliability at each level of multilevel data because single-level reliability is "difficult to interpret when reliability is not identical across levels" by averaging "across levels of measurement" (p. 77). Notably, level-specific composite reliability can be easily estimated using the parameter estimates from a multilevel CFA (Geldhof et al., 2014).

Geldhof et al. (2014) introduced multilevel versions of three commonly used reliability estimates: multilevel alpha, multilevel composite reliability, and multilevel maximal reliability. The formula of alpha (α; Cronbach, 1951) for single-level measurement can be directly used to estimate level-specific multilevel alpha:

$$\alpha = \frac{k^2 \bar{\sigma}_{ij}}{\sigma_X^2}$$

where k is the number of items in a scale, $\bar{\sigma}_{ij}$ is the average interim covariance within a scale that can be computed by averaging all unique covariances of observed indicators, and σ_X^2 is the variance of the scale scores that can be obtained as the sum of variances and twice each unique covariance of all observed indicators. With two-level data, the within- and between-level alphas can be computed by specifying saturated models at both levels and obtaining the variances and covariances of indicators at within and between levels, respectively.

Similarly, multilevel composite reliability (ω) and multilevel maximal reliability (H) can be computed with the formulae of ω and H for single-level measurement, respectively.

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii}};$$

$$H = \frac{\sum_{i=1}^{k} \frac{\ell_i^2}{1-\ell_i^2}}{1 + \sum_{i=1}^{k} \frac{\ell_i^2}{1-\ell_i^2}}, \quad \frac{\ell_i^2}{1-\ell_i^2} \approx \frac{\lambda_i^2}{\theta_{ii}}$$

where λ_i, θ_{ii}, and ℓ_i^2 are the factor loading, residual variance, and squared standardized factor loading of indicator i of a single factor model, respectively. To estimate level-specific ω and H, the factor loadings and residual variances of the corresponding level should be used. According to the simulation results of Geldhof et al. (2014), multilevel H is least recommended.

It is suggested to report level-specific reliability because a single-level reliability estimate with multilevel measurement conflates the within- and between-level reliability estimates and is not interpretable when the within- and between-level reliability estimates are heterogeneous. In addition, when reporting scale reliability, researchers should include information on the type and level of reliability (i.e., what type of reliability at which level). Otherwise, the meaning of the reported reliability is vague to readers.

Model fit evaluation

The fitted MFA models are typically evaluated with fit indices commonly used for single-level factor analysis such as the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). However, Ryu and West (2009) pointed out that fit indices available in most SEM software programs (except SRMR in M*plus* to the best of our knowledge) assess overall model fit and are not appropriate to evaluate model fit at each level, especially for Level 2 because the overall model fit is dominated by Level 1 fit information. Hsu, Kwok, Lin, and Acosta (2015) recently confirmed via a simulation study that CFI, Tucker-Lewis index (TLI), and RMSEA were not sensitive to the Level 2 model misspecification whereas the standardized root mean square residual (SRMR) between could detect such misspecification reasonably well. Thus, evaluating level-specific model fit is important particularly when the construct of interest is at the second level. Residuals (i.e., the element-wise difference between observed and model-implied covariance matrices) are also analyzed to evaluate local model misfit at each level. It should be kept in mind that model fit criteria such as CFI $\geq$ .95 and RMSEA $\leq$ .05 (Browne & Cudeck, 1993; Hu & Bentler, 1999) were proposed in the context of single-level CFA. Thus, there is a need for caution when employing these cutoff criteria for MFA. In terms of reporting practices, the methods used to evaluate MFA models (e.g., model fit criteria) should be justified and transparent to the readers.

Method

Identification of multilevel factor analysis applications

Following the reporting guidelines listed in Table 2, we explored the reporting practices of MFA in the current literature. Multilevel factor analysis applications were obtained from journals published between 1994 and 2014. The following databases were used to identify the sample of articles: WilsonOmnifile, PsychINFO, Google Scholar, MEDLINE, and Education (full text). Keywords used in the search were multilevel confirmatory factor analysis, multilevel factor analysis, and multilevel exploratory factor analysis. The search terms initially returned 216

articles. From the initial pool, 72 articles were excluded because the keywords merely appeared in the texts or references of the articles but MFA was not actually used in the study. Reference lists for the remaining 144 articles were further checked for additional applications. We sorted the articles into six categories: (a) Factor analysis was conducted with adjusted standard errors and chi-square tests to take into account the nonindependence of the data from multistage sampling (Muthén & Satorra, 1995), for example, the single-level approach using the TYPE = COMPLEX option in M*plus* ($n = 2$); (b) MFA was used as part of preliminary analyses only and was not the major focus of the study ($n = 41$); (c) MFA was the primary analysis and was used to address major research questions ($n = 72$); (d) didactic illustrations of MFA with real data applications ($n = 23$); (e) conceptual discussions of MFA without real data applications ($n = 5$); and (f) use of pooled within-covariance matrix ($n = 1$). We focused on the 72 articles with MFA as the primary analysis (category c) because these studies provided more complete descriptions of the procedures and results of the analyses compared to the other categories of studies. Conceptual articles on MFA were used to inform our coding protocol but were not used in the review.

Coding protocol

The coding instrument was developed and refined in stages and was based on several sources including a review of the (a) literature on technical and methodological issues related to MFA, (b) journal articles discussing statistical reporting practices, including the Standards for Reporting on Empirical Social Science Research in AERA Publications, and (c) discussions among our research group members. The coding protocol had 72 specific items addressing five major areas: (a) characteristics of the MFA application (type of nested structure, construct measured, number of items, number of dimensions, response scale); (b) purpose (e.g., measurement validation); (c) data source, including the number of cases at Level 1 and Level 2; (d) statistical approach, including software used and estimation methods (e.g., maximum likelihood); and (e) results reported (e.g., intraclass correlations for indicators and latent variables, standardized and unstandardized loadings, fit indices).

To evaluate the reliability of the coding procedures, all members of the research team independently coded five articles. Agreement ranged from 86% to 99% on the coded elements ($M = 94\%$). Kappa for multiple raters was computed using Stata version 13.1 (2013). Kappas for five articles were .79, .91, .92, .93, and .97, and the overall kappa was .90. Disagreements were discussed, and sections of the articles were reread to resolve disagreements.

The results presented in the next section are based on the 72 MFA applications that were reviewed using the 72-item coding protocol. The reviewed articles are listed in References and noted with an asterisk.

Results

The characteristics of the 72 MFA reviewed applications (e.g., type of nested data structure; exploratory or confirmatory) are presented in Table 3. The reporting practices of the applications of MFA (e.g., the proportions of applications that discussed the rationale of MFA) are summarized in Table 4. Descriptive statistics (i.e., mean, standard deviation, minimum, and maximum) of the reported MFA results and parameter estimates from the 72 reviewed articles are presented in Table 5. The reported values in Table 5 are expected to be useful for methodologists implementing simulation studies as well as for applied researchers to design their studies. Because the tables are self-explanatory, the following section is devoted to some points not shown in the tables and some findings warranting further attention.

Characteristics of the multilevel factor analyses

The 72 MFA applications were distributed across a wide range of U.S. (e.g., *Educational and Psychological Measurement*) and international journals (e.g., *Scandinavian Journal of Educational Research*) representing diverse fields of study (e.g., education, health, psychology, exercise science). The variability in the applications is reflected

Table 3. Characteristics of multilevel factor analysis ($N = 72$).

Characteristics	Proportion	Characteristics	Proportion
Data structure: Individuals within groups	.82	Type of FA: MCFA	.85
Repeated measure	.15	MEFA	.11
Other	.03	Both	.04
Construct reference[a]: Level 1	.42	Higher-order examined: Yes[c]	.13
Level 2	.58	Structural regression model examined: Yes[c]	.29
Scale: Continuous	.15	Nuisance factors modeled: Yes[c]	.26
Categorical	.78	Nuisance factors[d]: Correlated residual	.84
Missing	.07	Method factor	.16
If categorical, treated as continuous[b]: Yes[c]	.88	Purpose: Construct validation	.82
Nonexperiemental: Yes[c]	.89	Other	.18
		Secondary data: Yes[c]	.43

Note. FA = factor analysis; MCFA = multilevel confirmatory factor analysis; MEFA = multilevel exploratory factor analysis. [a] Construct reference means whether a construct refers to the characteristics of Level 1 units or Level 2 units (e.g., students' self-esteem vs. teacher control behaviors when students are nested within classrooms). [b] Proportion with applicable cases only ($N = 56$). [c] The remaining proportion is for the category No. [d] Proportion with applicable cases only ($N = 19$).

Table 4. Reporting practice of multilevel factor analysis: Proportion of articles that reported the following items ($N = 72$).

General	Proportion	Data	Proportion
Diagram	.42	Level 1 sample size	.97
Rationale of MFA	.96	Level 2 sample size	.97
Univarate normality	.24	Cluster size	.54
Multivariate normality	.04	Level 1 random sampling	.10
Analysis		Level 2 random sampling	.17
Covariance/correlation matrix	.46	Level 1 *N* justification	.07
Estimation method	.61	Level 2 *N* justification	.10
Convergence problems	.11	Level 1 power analysis	.01
Non-positive definite	.03	Level 2 power analysis	.03
Negative residual variance	.11	Level 1 complete data[a]	.13
Multiple models constructed	.44	Level 2 complete data[b]	.26
Software program used	.93	Level 1 missing data[a]	.49
Muthén's five-step followed	.11	Level 2 missing data[b]	.14
Parceling used	.16	Level 1 missing method if applicable	.86
Cross-level invariance tested	.08	Level 2 missing method if applicable	.80
Reliability Not reported	.46	Sampling weight applied if applicable	.00
Level not specified	.16	Item ICC	.64
At both levels	.17	Factor ICC	.24
Level 1	.17	Design effect	.07
Level 2	.04		

Note. MFA = multilevel factor analysis; ICC = intraclass correlation. [a] Proportion of missing = .38. [b] Proportion of missing = .60.

in the constructs that were examined (e.g., working alliance, neighborhood walkability, pain experience, collective teacher efficacy, intrinsic motivation in exercise, quality of health care service, coaching competency, product creativity, phonological awareness). Some of these constructs clearly represent characteristics of organizations or groups (e.g., community risk factors, leadership, work–family culture) while other constructs are characteristics of individuals (e.g., self-esteem of students, personality traits, and social attitudes) nested or clustered within organizations. Fifty-eight percent of the constructs ($n = 42$) referred to characteristics of Level 2 units and were modeled at the second level (e.g., teacher control behaviors perceived by students). These cases include repeated measures applications, such as diary entries nested within individuals, and thus in some cases the constructs at the second level represented individual characteristics.

The most common multilevel structure (60 out of 72; 83%) consisted of individual units at Level 1 (e.g., students) nested within groups at Level 2 (e.g., classes). We refer to this type of multilevel structure as an organizational structure. We classified a dyad study that used individuals (Level 1) within married couples (Level 2) as an organizational structure. About 15% of the studies (i.e., 11 articles) conducted MFA using repeated measures (Level 1) within persons (Level 2), which included daily diary studies.

Of the 67 articles (86%) that reported the type of observed variables (i.e., continuous or ordered categorical), 11 indicated that the data were continuous. On the other hand, 56 of the articles used questionnaires as their source of data, with ordered categorical indicators. Five-point scales were the most frequently used response format in the questionnaires ($n = 24$), followed by four-point scales ($n = 10$). In the majority of these articles using Likert-type scales, the ordered-categorical variables were treated as continuous ($n = 49$). When the estimation method was reported in these applications, the method (i.e., maximum likelihood [ML]) was appropriate for continuous data. This practice is not unexpected given that the measurement literature has suggested that ML performs reasonably well with five or more response categories whereas weighted least squares with mean and variance adjustment (WLSMV) is recommended with two or three response categories (Beauducel & Herzberg, 2006; Dolan, 1994). Nine studies that used a scale with fewer than five response categories used ML or robust maximum likelihood (MLR). Although these cases were not common, using an optimal estimation method in line with the type of data needs more attention in applications of MFA. Methodological endeavors to provide practical guidelines in the use of ordinal measures in the context of MFA are also called for.

Question wording in the data collection instruments used in the studies varied in terms of the referent used in the items. For example, Holfve-Sabel and Gustafsson (2005) used two types of questions related to students' interest in schooling—"I think the work in lessons is fun" and "Work at school is good and has variety"—in which the first question directly referenced the individual while the second question referenced the school. Holfve-Sabel and Gustafsson noted that the reference used in the item (e.g., individual) is a factor that can influence the variance at each level (within-group level) of the analysis.

Purposes of the multilevel factor analyses

The majority of the MFA applications (82%; $n = 59$) had a psychometric focus designed to address measurement validity issues. For example, Mathisen, Torsheim, and Einarsen (2006) conducted a two-level CFA of the Team Climate Inventory to test "the validity of climate scores as a measure of team or organizational climate as against merely a measure of the idiosyncratic perceptions of individual employees" (p. 31). In addition to the psychometric focus of most of the articles, 13 articles (18%) employed MFA for different purposes. One purpose that was evident in an article was to answer substantive

Table 5. Summary of data, fit statistics and parameter estimates of multilevel factor analysis ($N = 72$).

Variable		*N*	Mean	*SD*	Minimum	Maximum
Number of items		71	20.31	14.42	4	70
Number of response categories if applicable		55	5.47	3.48	2	28
Sample size Level 1		70	4548.46	8240.66	122	50513
Level 2		70	176.61	356.96	15	2267
Cluster size Minimum		31	10.87	18.84	1	80
Maximum		31	62.07	130.15	2	710
Mean		39	26.71	53.07	2	295.76
Number of factors Level 1		72	3.94	2.63	0	14
Level 2		68	3.34	2.64	0	14
Number of indicators per factor	Level 1 minimum	69	3.86	2.09	1	10
	Level 1 maximum	69	6.54	3.52	2	20
	Level 2 minimum	62	5.06	4.33	1	27
	Level 2 maximum	62	8.08	5.89	2	35
SRMR Within		36	.04	.02	.01	.11
SRMR Between		37	.08	.06	.01	.24
RMSEA		58	.04	.02	.00	.11
CFI		52	.95	.05	.77	1.00
TLI/NNFI		24	.96	.04	.85	1.00
Item ICC minimum		46	.13	.14	.01	.65
Maximum		46	.34	.18	.04	.93
Factor ICC minimum		17	.22	.21	.01	.66
Maximum		17	.36	.23	.11	.93
Design effect minimum		5	1.80	0.32	1.47	2.19
Maximum		5	5.78	2.26	3.38	9.22
Factor loading[b]	Level 1 minimum	54	0.41	0.20	0.02	0.82
	Level 1 maximum	54	0.83	0.10	0.52	0.99
	Level 2 minimum	47	0.47	0.25	0.03	0.91
	Level 2 maximum	47	0.94	0.19	0.34	1.62
Residual variance[c]	Level 1 minimum	11	.24	.14	.02	.53
	Level 1 maximum	11	.77	.19	.45	.99
	Level 2 minimum	8	.05	.06	.00	.17
	Level 2 maximum	8	.73	.25	.20	.95
Factor correlation[b]	Level 1 minimum	48	.35	.27	.00	.89
	Level 1 maximum	48	.68	.21	.23	.99
	Level 2 minimum	34	.42	.28	.00	.94
	Level 2 maximum	34	.72	.27	.00	1.00
Reliability	Level 1 minimum	24	.76	.19	.02	.93
	Level 1 maximum	24	.86	.07	.73	.96
	Level 2 minimum	15	.73	.21	.39	.92
	Level 2 maximum	15	.93	.06	.81	1.00

Note. *SD* = standard deviation; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI/NNFI = Tucker-Lewis index/nonnormed fit index; ICC = intraclass correlation. [a] Saturated model. [b] Absolute values of standardized factor loadings. [c] Standardized residual variance. One article reported root mean residual (RMR; .067); goodness-of-fit index (GFI; mean = .97) from 4 articles; normed fit index (NFI; mean .99) from 2 articles; chi-square/df ratio (mean 3.56) from 5 articles. Because multilevel exploratory factor analysis (MEFA) and multilevel confirmatory factor analysis (MCFA) did not show noticeable differences in the summary statistics, the results include both.

questions, such as the relative influence of schools on children's outcomes. Kuhlemeier, van den Bergh, and Rijlaarsdam (2002) used a multilevel exploratory factor analysis model similar to the three-level HLM of Raudenbush et al. (1991) to partition the variance in speaking and writing apprehension scores to determine whether there was more variability within schools or between schools. "The general finding seems to be that between school differences in the affective domain are much smaller than they are in the cognitive domain, indicating that schools are exerting more influence on students' achievements than on their attitudes and anxieties" (Kuhlemeier et al., 2002, p. 470).

In total, 69 applications (96%) justified the use of multilevel factor analysis to some degree. However, there was a wide variability in the complexity of the rationales for using MFA. For example, Breevaart, Bakker, Demerouti, and Hetland (2012) justified the use of multilevel analysis on the basis of high ICCs (.36–.55) observed in their data; Whitton and Fletcher (2014) used MFA because the construct of their study (e.g., group cohesion) is considered a group-level construct. Several researchers addressed the issue of aggregating individuals' scores up to a higher level, noting that it was important to establish the psychometric soundness of the measures "at the level they purport to inform" (Martin, Malmberg, & Liem, 2010, p. 976). D'Haenens, Van Damme, and Onghena (2010) observed that "aggregation of the teacher-level factors into the school level assumes functional correspondence between the latent dimensions at the teacher level and the measures for school process variables" (p. 157). A sentiment that was shared by many was that the "hierarchical nature of most educational and psychological data renders the examination of factor structure of the data

at different levels a necessary, if not compulsory, step before researchers are justified to conduct analyses at different levels of the data structure" (Martin et al., 2010, p. 977). Some studies discussed consequences resulting from single-level approaches to multilevel data in support of the use of MFA (e.g., Diya, Li, van den Heede, Sermeus, & Lesaffre, 2014). Those consequences included underestimated standard errors due to the violation of the independence assumption and improper inferences of the findings from one level to another, such as the atomistic fallacy that incorrectly extrapolates the effects found at the individual level to the cluster level and the ecological fallacy that falsely infers the effects at Level 1 on the basis of the results at Level 2.

Although there was great sensitivity to issues surrounding multilevel data structures and the effects of clustered data on standard errors, many studies did not provide extensive theoretical explanations of the meaning of their constructs within a multilevel framework. For example, in one application the researchers analyzed the dimensionality of students' self-concept at the individual and class levels but did not explain what self-esteem would mean at the class or macro level. Similarly, in an application involving a multilevel confirmatory factor analysis of students' perceptions of homework quality, the meaning of this classroom-level construct was not explained. These results suggest that in some cases the conceptualization of constructs has lagged behind the statistical advances evident in multilevel analyses. The lag in theoretical frameworks to support multilevel constructs poses a barrier to the construct validation process.

Data

MFA was generally conducted with an overall large sample. The median Level 1 sample size was 1,505.00 ($M = 4{,}548.46$), and the median Level 2 sample size was 83.50 ($M = 176.61$). Detailed information related to the data used to estimate the MFAs was generally limited. Only a small number of studies discussed sampling methods. None of the studies we reviewed applied sampling weights in data analysis. Researchers seldom reported that power analyses were conducted to estimate the sample size before data collection (one study to determine Level 1 sample size and two studies for Level 2 sample size). Forty-four studies (38%) did not explicitly mention whether the data were complete in terms of missingness at Level 1 or at Level 2.

Model specification

Thirty of the 72 MFA applications (42%) included one or more diagrams to represent the models. Only one application included a diagram like Figure 1, part (a). Fourteen applications followed a format similar to Figure 1, part (b) or (c) (two of them similar to Figure 1, part (c)). Seven applications provided diagrams similar to Figure 1, part (d); Figure 1, part (e) was found in five applications. The other three had either a single-level diagram or a diagram of a structural model (i.e., no measurement model). As discussed in the reporting guidelines, a schematic diagram similar to Figure 1, part (a) is recommended. Parts (d) and (e) of Figure 1 provide less information about the statistical model.

The major application of multilevel factor analysis was confirmatory. Of 72 articles reviewed, 61 (85%) conducted multilevel confirmatory factor analysis (MCFA) whereas 8 (11%) employed multilevel exploratory factor analysis (MEFA) to investigate the level-specific factor structures. Three studies (4%) conducted both MEFA and MCFA. It is interesting to note that 19 studies (27%) explicitly modeled a certain type of unintended effect or nuisance factor. In addition to the primary constructs that a scale is developed to measure, nuisance factors are sometimes present for different reasons, such as wording similarity in some items (e.g., negatively worded items; generically called method effects) or the presence of unexplained constructs. Two strategies were observed in those 19 applications: (a) allowing correlations between residuals ($n = 16$) or (b) directly modeling method effects as factors using multitrait multimethod (MTMM) models ($n = 3$). Researchers often included correlated residuals of observed variables mostly to improve model fit, but did not provide a theoretical rationale or explain the meaning of the correlated residuals. Because the error correlation may indicate the presence of nuisance factors or multidimensionality of a measure, modifying a model using error correlations should be guided by theoretical support, especially when MFA is carried out for construct development and validation.

At both Level 1 and Level 2, the number of factors specified or identified in the MFA ranged from 0 to 14. We coded the number of factors of a saturated model as 0. In one study, the within-covariance matrix was not analyzed and only the Level 2 factor structure was modeled using the between-covariance matrix because the construct (namely, teachers' control behaviors perceived by students) was a Level 2 construct. In other studies ($n = 3$), the factor structure at Level 2 was not of interest, and thus a saturated model was specified at Level 2 to take into account the variance between clusters. In 50 studies (69%), the numbers of factors at Level 1 and Level 2 were identical. When the numbers of factors at Level 1 and Level 2 were different ($n = 22$ studies; 31%), all reported a smaller number of factors at Level 2.

Results

Model fit indices and parameter estimates reported in MEFA versus MCFA, and in the studies using repeated measures versus organizational-type data were not notably different. Thus, we present the results of the total sample of studies in Table 5. ICCs are indices of data dependency and further indicate whether there is substantial variability across groups that needs to be modeled with between-level latent factors in the construct validation process. Especially, for a shared cluster-level construct, large ICCs of item responses provide evidence of within-group agreement and one source of support for the construct validity of a shared construct. However, not all studies provided ICCs. Forty-six articles (65%) reported item ICCs to document the lack of independence in the data and support the need for multilevel analysis (over a single-level analysis), and 17 articles (24%) reported factor ICCs. We observed a wide variability of ICCs from near zero to close to 1.00 across studies. In a dyad study with husband and wife, the item ICCs ranged from .65 to .93. The factor ICCs of this study were also high and ranged between .66 and .93. Without this dyad case, for the organizational-type multilevel structure (individuals within groups, $n = 38$) the minimum ICCs were between .01 and .31 ($M = .09$, $SD = .07$) and the maximum ICCs were between .04 and .70 ($M = .30$, $SD = .14$); for the repeated measures multilevel structure ($n = 7$), the minimum ICCs were between .07 and .64 ($M = .29$, $SD = .18$) and the maximum ICCs were between .19 and .77 ($M = .49$, $SD = .18$). The factor ICCs were in a similar range with a slightly higher maximum (.82).

As shown in Table 5, when reported, the alternative fit indices generally supported good model fit. We found only one article with the RMSEA $> .08$, five articles with CFI $< .90$, and one article with an SRMR within $> .08$. However, SRMR between values were higher, with the mean of .08 ($SD = .06$, minimum $= .01$, maximum $= .24$) and greater than .08 in 13 cases. The overall promising values of alternative fit indices should be interpreted with caution because a number of applications did not present these fit indices. In addition, eight studies did not specify the level of SRMR.

Cross-level invariance was seldom evaluated in the reviewed applications. Six applications (8%) tested the equality of factor loadings across levels using likelihood ratio tests, and only two achieved cross-level invariance. In terms of reliability of a measure, a total of 39 studies (54%) presented reliability. Twelve studies reported Level 1 reliability, three studies reported Level 2 reliability, and 12 studies presented reliability at both levels. In another 12 studies, reliability was reported but the level at which it was computed was not specified.

Discussion

The results of this systematic review were based on 72 MFA applications across a variety of disciplines. We provide a report on how these techniques are being used and what results are being presented. These techniques are complex and relatively new, and thus the results of these studies as well as the reporting guidelines of MFA can provide guidance to applied researchers interested in expanding their approaches to psychometric analyses and construct validation. This guidance is particularly important given the range and number of multilevel constructs in areas such as organizational psychology and educational assessment. Although the reviewed articles were generally consistent with the guidelines designed to increase transparency in research reporting, the results suggest that there are areas in need of improvement. These areas include the reporting of tests of distributional assumptions, discussion of missing data, analysis of statistical power given the number of Level 1 and Level 2 units, the degree of nesting reflected in the ICC, identification of the sampling methods (e.g., probability vs. nonprobability) used at each level, discussion of measures of model fit in a multilevel context, and more detailed descriptions of "considerations during the data analysis that might compromise the validity of the statistical analyses" (American Educational Research Association, 2006, p. 37). These considerations include estimation problems, lack of convergence, improper solutions, and consequences of ignoring a level in the analysis. In the present study, only eight studies mentioned convergence problems or the need to set residual variances at Level 2 to zero. Several studies implied a possible higher level of units and explicitly raised questions about ignoring a level in the analysis. An example related to this issue involved a three-level data structure (i.e., respondents within hospital units within hospitals) in which the researchers implemented two separate analyses—individuals within hospital units and individuals within hospitals.

Although we did not code the type of constructs using the taxonomy of Stapleton et al. (2016) introduced earlier, it appears that many of the reviewed applications are not neatly labeled with such nomenclature and the Level 2 constructs in reality are often "fuzzy composition" (Bliese, 2000, p. 369) falling somewhere between shared and configural constructs. At minimum, we found that in 42 articles (58%), the construct of interest was at the organizational level (e.g., group cohesion and work climate) whereas 30 studies (42%) dealt with the construct at the individual level (e.g., self-efficacy and cognitive ability). However, explicit discussions of how researchers conceptualize the constructs in their studies and how the constructs (or models) are specified, estimated, and interpreted at each level are lacking. Such information

is critical in the construct validation of multilevel measures and should be fully discussed in the research article. As observed in this review, usually the rationale of MFA was not extensively discussed. Often the researchers justified the use of MFA mainly by the presence of a nested data structure. Given the complexity of multilevel factor structure, more in-depth discussions of hypothesized constructs should be included. Moreover, it is essential for the purpose of construct validation in multilevel data to interpret the meaning of constructs identified at each level.

It is somewhat surprising that only 64% of the reviewed studies reported the item-level ICCs, with fewer studies reporting the factor ICCs. The ICC is a ratio of Level 2 variance to the total variance and is an index of data dependency. Beyond serving as an index of data dependency, high ICCs indicate considerable variability at Level 2 and thus the need to model latent factors at Level 2, especially shared cluster constructs in the multilevel FA. Similarly, a variant of the ICC that takes into account cluster size (i.e., the number of Level 1 units per cluster) is interpreted as a reliability index for Level 2 shared construct (Bliese, 2000; Lüdtke et al., 2008) because individual responses of the shared construct are expected to be isomorphic and large variability across individuals at Level 1, which results in low ICCs, can indicate low reliability of the shared construct. Without information on ICCs, the sources of variability are not clear to readers. Hence, the use of the multilevel approach to construct validation cannot be fully justified, and the measurement quality of Level 2 shared constructs cannot be properly evaluated. As illustrated in Muthén's five-step approach (computation of the ICC at Step 2), evaluating and reporting a measure of data dependency such as item-level ICCs should be a standard reporting practice. It should be noted that item-level ICCs are readily obtainable as a part of the M*plus* default output.

In our review, only six applications tested cross-level invariance, and only two of these applications achieved cross-level invariance. In relation to the reports of factor ICCs, two studies that presented factor ICCs conducted cross-level invariance tests, and one of them met the cross-level invariance assumption in computing the factor ICCs. As demonstrated in this review, cross-level invariance is either not recognized as an important condition for certain MFA models or ignored among applied researchers conducing MFA. This review also showed that the violation of cross-level invariance is not uncommon (4 of 6 studies). When cross-level invariance is violated, the interpretation of MFA results is often challenging, but there is limited discussion focusing on the lack of cross-level invariance in either the applied or methodological literature of MFA. More attention to cross-level invariance among applied researchers as well as practical suggestions and guidelines from methodologists in cases where there is a lack of cross-level invariance are warranted.

Although the emphasis on level-specific reliability appeared as early as in Muthén's (1991) illustration of MFA, only 12 studies (17%) reported level-specific reliability estimates (i.e., reliability estimates at each level). The researchers of these 12 studies, which were published as early as 1997 and as late as 2014, seemed aware of the issues of single-level reliability estimates for the constructs specified at multiple levels. However, many studies still reported single-level reliability estimates or did not specify the level at which reliability was calculated. More troubling from a measurement perspective is not reporting reliability in almost half of the reviewed studies (46%) although most studies in this review purported to validate constructs of interest. Because reliability is essential for a measure and considered as a prerequisite for measurement validity, reporting reliability is vital in the process of construct validation.

It appears that the choice of model fit indices heavily depended on the software program used for MFA. Researchers tend to evaluate and present the fit indices available in the software program. Given the dominance of M*plus* as a choice of software for MFA, it is not surprising that most studies reported RMSEA, CFI, TLI, and SRMR, which are produced by default in the M*plus* output. Although there are dozens of fit indices developed for the evaluation of model fit in the SEM literature, reporting only software default fit indices is not of great concern because many fit indices are highly correlated with each other and do not necessarily provide unique information of model fit (e.g., CFI, Incremental Fit Index, Relative Noncentrality Index, Gamma hat, and McDonald's Noncentrality Index; Cheung & Rensvold, 2002). However, we did not find any study reporting level-specific fit indices other than the SRMR within and between (with the exception of four studies that constructed a saturated model at one of the levels). Beyond the criticism of researchers' excessive reliance on model fit cutoffs developed by several researchers (e.g., Browne & Cudeck, 1993; Hu & Bentler, 1999), it should be noted that these fit criteria (e.g., CFI $\geq$ .95) were developed for a single-level SEM. Considering researchers' high reliance on the default output in evaluating and reporting model fit, methodologists need to assess the behaviors of common model fit indices in the context of MFA and consider developing level-specific fit indices for model evaluation. These fit measures could then be included in MFA software programs.

Overall, sampling methods were rarely described in the reviewed applications (e.g., complex two-stage sampling with random sampling of Level 2 units and then random sampling of Level 1 units within selected Level 2 units). In addition, sampling weights were not applied in any of the

reviewed applications. It is not known in how many cases sampling weights were available but not applied. However, if applicable, sampling weights should be applied to make inferences to the population. Asparouhov and Muthén (2006) stated that "if the sampling weights are ignored at either level, the parameter estimates can be substantially biased" (p. 2718) due to unequal probabilities of selection in sampling. In many multilevel software programs, sampling weights can be applied with a simple statement or an option.

Along with the increase in the applications of MFA, Monte Carlo simulation studies investigating the behaviors of MFA in a variety of circumstances have increased in recent years (e.g., Can, van de Schoot, & Hox, 2015; Hsu et al., 2015; Jak et al., 2013; Kim & Cao, 2015). Boomsma (2013) provided general guidelines for conducting Monte Carlo studies in SEM and emphasized the importance of designing these studies such that their findings generalize to real world research. Thus, when choosing a population model and population parameters, it is recommended that simulation researchers "review structural equation model applications across a large number of journals in several areas of research to which they would like to generalize the subsequent results" (Paxton, Curran, Bollen, Kirby, & Chen, 2001, p. 291). As Boomsma acknowledged, this type of thorough review of a large pool of literature is tedious and time consuming, and it is not unusual for simulation researchers to justify the selected model and parameters on the basis of a minireview of literature or on previous simulation studies. Hence, the results of our systematic review, which included the values of parameter estimates as well as details of research settings, will be particularly useful for simulation researchers designing Monte Carlo studies on MFA.

The results of the present study should be viewed with the following delimitations and limitations in mind. We searched a limited number of databases and thus applications in some fields might be systematically excluded. We also focused on only one category of MFA applications—those applications in which MFA was the primary analysis used to address major research questions. Of note is that multilevel item response theory (IRT) applications based on frameworks such as those by Fox and Glas (2001) and hierarchical generalized linear modeling (e.g., Kamata, Bauer, & Miyazaki, 2008) were not explicitly included in the present review. However, because CFA with ordered-categorical variables is analogous to IRT (e.g., Wirth & Edwards, 2007), particularly when the maximum likelihood estimation is used, and MCFA is one approach to multilevel IRT (Kamata & Vaughn, 2011), some applications using categorical variables can also be considered as multilevel IRT applications.

In addition, our coding focused only on what was reported in the journal article. The absence of certain results in the published article may not necessarily mean that the researcher did not conduct analyses that would have produced these results. For example, a researcher may have conducted extensive analyses to evaluate multivariate normality but because of space limitations was not able to report these results. Although we achieved acceptable levels of reliability in coding 72 aspects of each article, several items that were coded required subtle judgments, which likely introduced measurement error into the coding process. In addition, it is important to note that the field of multilevel modeling is dynamic, with new issues and guidelines emerging on a regular basis. Our review concluded with articles published in 2014, and therefore future research will need to continuously reevaluate methodological practices and reporting guidelines and update the 72-item protocol to account for changes in the field. Notwithstanding these limitations, the significance of the current study is that it has raised awareness of a number of technical and reporting issues related to multilevel factor analysis and has provided direction to researchers interested in conducting simulation studies that reflect realistic design conditions. Through discussion of these issues informed by empirical research, applied researchers will be able to achieve the maximum benefits of these statistical methods.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant from any funding agency.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors thank anonymous reviewers for their comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the

authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

References

References marked with an asterisk indicate articles that were reviewed using the 72-item coding protocol.

*Allodi, M. W. (2002). A two-level analysis of classroom climate in relation to social context, group composition, and organization of special support. *Learning Environments Research*, *5*, 253–274.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40.

Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting, Seattle, WA, ASA section on Survey Research Methods*, 2718–2726.

*Bakali, J., Baldwin, S., & Lorentzen, S. (2009). Modeling group process constructs at three stages in group psychotherapy. *Psychotherapy Research*, *19*, 332–343.

Bandalos, D. L, & Finney, S.J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Muller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). New York, NY: Routledge Taylor & Francis.

*Barile, J., Darnell, A., Erickson, S., & Weaver, S. (2012). Multilevel measurement of dimensions of collaborative functioning in a network of collaboratives that promote child and family well-being. *American Journal of Community Psychology*, *49*, 270–282.

*Beadnell, B., Carlisle, S. K., Hoppe, M. J., Mariano, K. A., Wilsdon, A., Morrison, D. M., &... Higa, D. (2007). The Reliability and validity of a group-based measure of adolescents' friendship closeness. *Research on Social Work Practice*, *17*, 707–719.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 186–203.

Bentler, P. M. (2000-2008). *EQS: Structural equations program manual*. Encino, CA: Multivariate Software.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In: K. J. Klein, & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organization*. San Francisco, CA: Jossey-Bass.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 518–540.

*Boonen, T., Pinxten, M., Van Damme, J., & Onghena, P. (2014). Should schools be optimistic? An investigation of the association between academic optimism of schools and student achievement in primary education. *Educational Research and Evaluation*, *20*, 3–24.

*Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classroom. *Journal of Educational Psychological*, *98*, 170–181.

*Breevaart, K., Bakker, A. B., Demerouti, E., & Hetland, J. (2012). The measurement of state work engagement: A multilevel factor analytic study. *European Journal of Psychological Assessment*, *28*, 305–312.

*Brondino, M., Pasini, M., & Silva, S. A. (2013). Development and validation of an Integrated Organizational Safety Climate Questionnaire with multilevel confirmatory factor analysis. *Quality and Quantity*, *47*, 2191–2223.

*Brown, E., Hawkins, J., Arthur, M., Abbott, R., & Van Horn, M. (2008). Multilevel analysis of a measure of community prevention collaboration. *American Journal of Community Psychology*, *41*, 115–126.

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

*Byman, R., Lavonen, J., Juuti, K., & Meisalo, V. (2012). Motivational orientation in physics learning: A self-determination theory approach. *Journal of Baltic Science Education*, *12*, 379–392.

Can, S., Van de Schoot, R., Hox (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimation. *Educational and Psychological Measurement*, *75*(3) 406–427.

*Cerin, E., Leslie, E., Owen, N., & Bauman, A. (2008). An Australian version of the neighborhood environment walkability scale: Validity evidence. *Measurement in Physical Education & Exercise Science*, *12*, 31–51.

*Cerin, E., Saelens, B. E., Sallis, J. F., & Frank, L. D. (2006). Neighborhood environment walkability scale: Validity and development of a short form. *Medicine & Science in Sports & Exercise*, *38*, 1682–1691.

*Cerin, E., Conway, T., Saelens, B., Frank, L., & Sallis, J. (2009). Cross-validation of the factorial structure of the Neighborhood Environment Walkability Scale (NEWS) and its abbreviated form (NEWS-A). *International Journal of Behavioral Nutrition and Physical Activity*, *6*, 32–42.

*Chen, Y., & Johantgen, M. E. (2010). Magnet hospital attributes in European hospitals: A multilevel model of job satisfaction. *International Journal of Nursing Studies*, *47*, 1001–1012.

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.

*Cheung, M. W., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research: An illustration with social axioms. *Journal of Cross-Cultural Psychology*, *5*, 522–541.

*Church, A. T., Katigbak, M. S., Ching, C. M., Zhang, H., Shen, J., Arias, R. M., &... Alvarez, J. M. (2013). Within-individual variability in self-concepts and personality states: Applying density distribution and situation-behavior approaches across cultures. *Journal of Research in Personality*, *47*, 922–935.

*Coromina, L., & Coenders, G. (2006). Reliability and validity of egocentered network data collected via web. A meta-analysis of multilevel multitrait multimethod studies. *Social Networks*, *28*, 209–231.

*Coromina Soler, L., Coenders, G., & Kogovsek, T. (2004). Multilevel multitrait-multimethod model: Application to the measurement of egocentered social networks. *Metodološki zvezki*, *1*, 323–349.

*Corsi, D. J., Subramanian, S. V., McKee, M., Wei, L., Swaminathan, S., Lopez-Jaramillo, P., &... Schooling, C. M. (2012).

Environmental profile of a community's health (EPOCH): An ecometric assessment of measures of the community environment based on individual perception, *Plos One*, *7*, 1–7.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometika*, *16*, 297–334.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J., & Lee, R. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69–102.

*Dedrick, R. F., & Greenbaum, P. E. (2011). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders*, *19*, 27–40.

*den Brok, P., Bergen, T., Stahl, R. J., & Brekelmans, M. (2004). Students' perceptions of teacher control behaviors. *Learning and Instruction*, *14*, 425–443.

D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Constructing measures for school process variables: The potential of multilevel confirmatory factor analysis. *Quality & Quantity*, *46*, 155–188.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326.

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*, 225–241.

*Diya, L., Li, B., Heede, K., Sermeus, W., & Lesaffre, E. (2014). Multilevel factor analytic models for assessing the relationship between nurse-reported adverse events and patient safety. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*, 237–257.

*Doss, B. D., Atkins, D. C., & Christensen, A. (2003). Who's dragging their feet? Husbands and wives seeking marital therapy. *Journal of Marital & Family Therapy*, *29*, 165–177.

*Droseltis, O., & Vignoles, V. L. (2010). Towards an integrative model of place identification: Dimensionality and predictors of intrapersonal-level place preferences. *Journal of Environmental Psychology*, *30*, 23–34.

*Dyer, N., Sorra, J., Smith, S., Cleary, P., & Hays, R. (2012). Psychometric properties of the consumer assessment of healthcare providers and systems (CAHPS (R)) clinician and group adult visit survey. *Medical Care*, *50*, S28–S34.

*Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction*, *29*, 1–9.

Ferron, J., Hogarty, K., Dedrick, R., Hess, M., Niles, J., & Kromrey, J. (2008). Reporting results from multilevel analyses. In A. O'Connell & D. B. McCoach (Eds.). *Multilevel analysis of educational data* (pp. 391–426). Charlotte, NC: Information Age Publishing.

Fox, J.P., & Glas, C. A. W. (2001). Bayesian estimation of multilevel IRT modeling using Gibbs sampling. *Psychometrika*, *66*, 271–288.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*, 72–91.

*Goetz, T., Lüdtke, O., Nett, U. E., Keller, M. M., & Lipnevich, A. A. (2013). Characteristics of teaching and students' emotions in the classroom: Investigating differences across domains. *Contemporary Educational Psychology*, *38*, 383–394.

*Goldstein, J., & McCoach, D. B. (2011). The starting line: Developing a structure for teacher ratings of students' skills at kindergarten entry. *Early Childhood Research & Practice*, *13*(2). Retrieved from http://files.eric.ed.gov/fulltext/EJ956366.pdf.

Graves, S. L., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Review*, *24*, 84–94.

*Greenbaum, P., Wang, W., Boothroyd, R., Kutash, K., & Friedman, R. (2011). Multilevel confirmatory factor analysis of the Systems of Care Implementation Survey (SOCIS). *Journal of Behavioral Health Services & Research*, *38*, 303–326.

*Haller, C. S., Courvoisier, D. S., & Cropley, D. H. (2011). Perhaps there is accounting for taste: Evaluating the creativity of products. *Creativity Research Journal*, *23*, 99–109.

*Hammer, L. B., Kossek, E., Bodner, T., & Crain, T. (2013). Measurement development and validation of the Family Supportive Supervisor Behavior Short-Form (FSSB-SF). *Journal of Occupational Health Psychology*, *18*, 285–296.

*Hammer, L., Kossek, E., Yragui, N., Bodner, T., & Hanson, G. (2013). Development and validation of a multidimensional measure of Family Supportive Supervisor Behaviors (FSSB). *Journal of Management*, *35*, 837–856.

*Hansson, A. (2012). Instructional responsibility in mathematics education: Modeling classroom teaching using Swedish data. *Educational Studies in Mathematics*, *75*, 171–189.

*Harnqvist, K., Gustafsson, J., Muthen, B. O., & Nelson, G. (1994). Hierarchical models of ability at individual and class levels. *Intelligence*, *18*, 165–187.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*, 393–416.

*Holfve-Sabel, M., & Gustafsson, J. E. (2005). Attitudes towards school, teacher, and classmates at classroom and individual levels: An application of two-level confirmatory factor analysis. *Scandinavian Journal of Educational Research*, *49*, 187–202.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.

Hox, J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*, 157–174.

Hsu, H-Y., Kwok, O., Acosta, S., & Lin, J-H. (2015). Detecting misspecified multilevel SEMS using common fit indices: A Monte Carlo study. *Multivariate Behavioral Research*, *50*, 197–215.

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

*Hulpia, H, Devos, G., Rosseel, Y., & Vlerick, P. (2012). Dimensions of distributed leadership and the impact on teachers' organizational commitment: A study in secondary education. *Journal of Applied Social Psychology*, *7*, 1745–1784.

*Intanam, N., Wongwanich, S., & Lawthong, N. (2012). Development of a model for building professional learning communities in schools: Teachers' perspectives in Thai educational context. *Journal of Case Studies in Education*, *3*, 1–11.

Jackson, D. L., Gillaspy , & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6–23.

Jackson, D. L. (2010). Reporting results of latent growth modeling and multilevel modeling analyses: Some recommendations for rehabilitation. *Rehabilitation Psychology*, *55*, 272–285.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 265–282.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Charlotte, NC: Information Age Publishing.

Kamata, A., & Vaughn, B. K. (2011). Multilevel IRT modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 41–57). New York: Taylor and Francis Group.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*, 350–386.

Kim. E. S., & Cao, C. (2015). Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. *Multivariate Behavioral Research*, *50*, 436–456.

*Klangphahol, K., Traiwichitkhun, D., & Kanchanawasi, S. (2010). Applying multilevel confirmatory factor analysis techniques to perceived homework quality. *Research in Higher Education Journal*, *6*, 1–10.

*Kuhlemeier, H., van den Bergh, H., & Rijlaarsdam, G. (2002). The dimensionality of speaking and writing: A multilevel factor analysis of situational, task and school effects. *British Journal of Educational Psychology*, *72*, 467–482.

*Li, F., Duncan, T. E., Duncan, S. C., Harmer, P., & Acock, A. (1997). Latent variable modeling of multilevel intrinsic motivation data. *Measurement in Physical Education & Exercise Science*, *1*, 223–224.

*Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the Life Skills Profile–16. *Psychological Assessment*, *25*, 810–825.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229.

Mackinnon, S. P., Battista, S. R., Sherry, S. B., & Stewart, S. H. (2014)*. Perfectionistic self-presentation predicts social anxiety using daily diary methods. *Personality and Individual Differences*, *56*, 143–148.

*Martin, A. J., Malmberg, L.-E., & Liem, G. A. D. (2010). Multilevel motivation and engagement: Assessing construct validity across students and schools. *Educational and Psychological Measurement*, *70*, 973–989.

*Mathisen, G. E., Torsheim, T., & Einarsen, S. (2006). The team-level model of climate for innovation: A two-level confirmatory factor analysis. *Journal of Occupational and Organizational Psychology*, *79*, 23–35.

*Mauno, S., Kiuru, N., & Kinnunen, U. (2011). Relationships between work-family culture and work attitudes at both the individual and the departmental level. *Work and Stress*, *25*, 147–166.

McCoach, D. B. (2010). Hierarchical linear modeling. In G.R. Hancock & R. O. Muller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 93–114). New York, NY: Routledge Taylor & Francis.

*McCoach, D. B., & Colbert, R. D. (2010). Factors underlying the collective teacher efficacy scale and their mediating role in the effect of socioeconomic status on academic achievement at the school level. *Measurement and Evaluation in Counseling and Development*, *43*, 31–47.

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

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analysis. *Psychological Methods*, *7*, 68–82.

*Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading*, *9*, 85–116.

*Merz, E. L., & Roesch, S. C. (2011). Modeling trait and state variation using multilevel factor analysis with PANAS daily diary data. *Journal of Research in Personality*, *45*, 2–9.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.8 for Windows [Computer software]*. Skokie, IL: Scientific Software International, Inc.

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology*, *32*, 83–104.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376–398.

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. Marsden (Ed.), *Sociological methodology* (pp. 216–316). Washington, DC: American Sociological Association.

*Myers, N. D., Chase, M. A., Beauchamp, M. R., Jackson, B. (2010). Athletes' perceptions of coaching competency scale II-high school teams. *Educational and Psychological Measurement*, *70*, 477–494.

*Nixdorf, D. R., John, M. T., Wall, M. M., Friction, J. R., & Schiffman, E. L. (2010). Psychometric properties of the modified Symptom Severity Index (SSI). *Journal of Oral Rehabilitation*, *37*, 11–20.

*O'Malley, A. J., Zaslavsky, A. M., Hays, R. D., Hepner, K. A., Keller, S., & Cleary P. D. (2005). Exploratory factor analysis of the CAHPS® hospital pilot survey responses across and within medical, surgical, and obstetric services. *Health Service Research*, *40*, 2078–2095.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 287–312.

*Peters, C., Algina, J., Smith, S. W., & Daunic, A. P. (2012). Factorial validity of the behavior rating inventory of executive

function (BRIEF)-Teacher form. *Child Neuropsychology*, *18*, 168–181.

*Radtke, T., Scholz, U., Keller, R., Knäuper, B., & Hornung, R. (2011). Smoking-specific compensatory health beliefs and the readiness to stop smoking in adolescents. *British Journal of Health Psychology*, *16*, 610–625.

Rasbash, J., Steele, F., Browne, W. J. & Goldstein, H. (2012). *A User's Guide to MLwiN, v2.26.* Bristolk, UK: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 295–330.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, *84*, 126–136.

*Rush, J., & Hofer, S. M. (2014). Differences in within-and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment*, *26*, 462–473.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 583–601.

Schreiber, J. B., & Griffin, B. W. (2004). Review of multilevel modeling and multilevel studies in *The Journal of Educational Research* (1992–2002). *The Journal of Educational Research*, *98*, 24–33.

*Schneider, S., Choi, S. W., Junghaenel, D. U., Schwartz, J. E., & Stone, A. A. (2013). Psychometric characteristics of daily diaries for the Patient-Reported Outcomes Measurement Information System (PROMIS®): A preliminary investigation. *Quality of Life Research*, *22*, 1859–1869.

*Sexton, J., Helmreich, R., Neilands, T., Rowan, K., Vella, K., Boyden, J., & Thomas, E. (2006). The Safety Attitudes Questionnaire: Psychometric properties, benchmarking data, and emerging research. *BMC Health Services Research*, *6*, 44–53.

*Sexton, J., Makary, M., Tersigni, A., Pryor, D., Hendrich, A., Thomas, E., &... Pronovost, P. (2006). Teamwork in the operating room: Frontline perspectives among hospitals and operating room personal. *Anesthesiology*, *105*, 877–884.

*Siriparp, T., Traiwichitkhun, D., & Kanjanawasee, S. (2012). Using multilevel confirmatory factor analysis to study student well-being in Thailand. *Journal of International Education Research*, *8*, 367–372.

Sirotnik, K. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement*, *17*, 245–282.

*Sorra, J., & Dyer, N. (2010). Multilevel psychometric properties of the AHRQ hospital survey on patient safety culture. *BMC Health Services Research*, *10*, 199–211.

*Stankov, L., & Lee, J. (2009). Dimensions of cultural differences: Pancultural, ETIC/EMIC, and ecological approaches. *Learning and Individual Differences*, *19*, 339–354.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520.

*Stawarczyk, D., Cassol, H., & D'Argembeau, A. (2013). Phenomenology of future-oriented mind-wandering episodes. *Frontiers in Psychology*, *4*, 425–436.

*Steele, F., & Goldstein, H. (2006). A multilevel factor model for mixed binary and ordinal indicators of women's status. *Sociological Methods & Research*, *35*, 137–153.

Thoemmes, F. J., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*, 90–118.

*Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272–296.

van de Vijver, F. J., & Poortinga, Y. H. (2002). Structural equivalence in multilevel research. *Journal of Cross-Cultural Psychology, 33*, 141–156.

*van Hemert, D. A., van de Vijver, F. J., Poortinga, Y. H., & Georgas, J. (2002). Structural and functional equivalence of the Eysenck Personality Questionnaire within and between countries. *Personality and Individual Differences*, *33*, 1229–1249.

*Van Horn, M., Hawkins, J., Arthur, M., & Catalano, R. (2007). Assessing community effects on adolescent substance use and delinquency. *Journal of Community Psychology*, *35*, 925–946.

*Wagner, W., Gollner, R, Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction*, *28*, 1–11.

*Whitton, S. M., & Fletcher, R. B. (2014). The group environment questionnaire: A multilevel confirmatory factor analysis. *Small Group Research*, *45*, 68–88.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future direction. *Psychological Methods*, *12*, 58–79.

*Wu, C. (2009). Factor analysis of the general self-efficacy scale and its relationship with individualism/collectivism among twenty-five countries: Application of multilevel confirmatory factor analysis. *Personality and Individual Differences*, *46*, 699–703.

*Zhang, N., & Wan, T. (2005). The measurement of nursing home quality: Multilevel confirmatory factor analysis of panel data. *Journal of Medical Systems*, *t*, 401–411.

*Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement*, *65*, 465–481.

Zumbo, B. D., & Forer, B. (2010, May). *A multilevel view of measurement validity: Some concepts and foundations*. Paper presented at the meeting of the American Educational Research Association, Denver, CO.

Zumbo, B. D., Liu, Y., Forer, B., & Shear, B. (2010, May). *National and international educational achievement testing: A case of multi-level validation*. Paper presented at the meeting of the American Educational Research Association, Denver, CO.

*Zúñiga, F., Schwappach, D., De Geest, S., & Schwendimann, R. (2013). Psychometric properties of the Swiss version of the nursing home survey on patient safety culture. *Safety Science*, *55*, 88–118.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice*, *12*, 127–140.