Psychological Methods
2014, Vol. 19, No. 1, 72–91

1082-989X/14/$12.00 DOI: 10.1037/a0032138

Reliability Estimation in a Multilevel Confirmatory Factor Analysis Framework

G. John Geldhof
Tufts University

Kristopher J. Preacher
Vanderbilt University

Michael J. Zyphur
University of Melbourne

Scales with varying degrees of measurement reliability are often used in the context of multistage sampling, where variance exists at multiple levels of analysis (e.g., individual and group). Because methodological guidance on assessing and reporting reliability at multiple levels of analysis is currently lacking, we discuss the importance of examining level-specific reliability. We present a simulation study and an applied example showing different methods for estimating multilevel reliability using multilevel confirmatory factor analysis and provide supporting Mplus program code. We conclude that (a) single-level estimates will not reflect a scale's actual reliability unless reliability is identical at each level of analysis, (b) 2-level alpha and composite reliability (omega) perform relatively well in most settings, (c) estimates of maximal reliability (H) were more biased when estimated using multilevel data than either alpha or omega, and (d) small cluster size can lead to overestimates of reliability at the between level of analysis. We also show that Monte Carlo confidence intervals and Bayesian credible intervals closely reflect the sampling distribution of reliability estimates under most conditions. We discuss the estimation of credible intervals using Mplus and provide R code for computing Monte Carlo confidence intervals.

Keywords: multilevel SEM, reliability, multilevel modeling, alpha, composite reliability

Supplemental materials: http://dx.doi.org/10.1037/a0032138.supp

Reliability has been defined alternatively as the squared correlation between true and observed scores (e.g., Lord & Novick, 1968, p. 61) or as the ratio of a scale's true score variance to its total variance (e.g., McDonald, 1999). These definitions are mathematically equivalent when the observed score's variance is positive (e.g., de Gruijter & van der Kamp, 2008; Raykov & Marcoulides, 2011), and both assume access to a scale's true score variance. Because true score variance is unknown and can only be estimated from observed data, most reliability estimates rely on the assumption that observed covariances necessarily represent true score variance.

This article was published Online First May 6, 2013.

G. John Geldhof, Institute for Applied Research in Youth Development, Tufts University; Kristopher J. Preacher, Department of Psychology and Human Development, Vanderbilt University; Michael J. Zyphur, Department of Management and Marketing, University of Melbourne, Melbourne, Victoria, Australia.

This research was supported in part by the University of Kansas's Center for Research Methods and Data Analysis and a grant from the John Templeton Foundation. We thank the University of Kansas Multilevel SEM workgroup for valuable input and feedback. We also thank Li Cai for comments on an earlier version of the manuscript.

Correspondence concerning this article should be addressed to G. John Geldhof, Institute for Applied Research in Youth Development, Lincoln Filene Building, Tufts University, Medford, MA 02155. E-mail: John.Geldhof@Tufts.edu

Reliability estimates are only as trustworthy as the information used to estimate them, however, and estimating reliability from data collected through multistage sampling necessarily confounds within- and between-cluster item variance (i.e., within-group variance and between-group variance). As such, multistage sampling may lead to biased reliability estimates when the assumption of independent residuals is violated (e.g., Snijders & Bosker, 1999).

Multistage sampling occurs when cases are randomly sampled from higher order units that are themselves sampled from a larger population of such units. For example, an education researcher might recruit several schools, select a sample of classrooms from each school, then obtain samples of students from each classroom (e.g., Connor et al., 2010). Multistage sampling results in hierarchically structured data (e.g., students nested within classrooms), making residuals dependent in the presence of between-cluster variation. Scores on key variables from children in a given classroom might be more alike than those of children in different classrooms, for instance. Ignoring hierarchical data structures can bias estimates of interitem relationships, likewise biasing reliability estimation for a desired level of analysis. Single-level reliability estimates therefore do not necessarily reflect true scale reliability at any single level of analysis.

The need to account for multilevel variability has been firmly established for hypothesis testing but has been largely ignored in the context of estimating a scale's reliability. It is commonly seen that researchers who appropriately use multilevel analysis to test primary hypotheses nevertheless report Cronbach's α as evidence

for acceptable levels of scale reliability, even though doing so implicitly assumes a single-level factor structure. This unfortunate status quo is not the fault of researchers, however, as extant methods for estimating reliability focus on a scale's total variability rather than the reliability of a scale at multiple levels of analysis (e.g., Cronbach, 1951; McDonald, 1999).

To give researchers a conceptual and operational foundation for understanding reliability at multiple levels of analysis, in the present article we describe how multilevel confirmatory factor analysis (MCFA) can be used to separately estimate reliability within and between clusters of a multilevel model. Whereas our logic can apply to any number of levels, we constrain our discussion to two-level models to facilitate presentation.

We first discuss common single-level reliability estimates and describe how each can be estimated within a confirmatory factor analysis (CFA) framework. We then address the dangers of misapplying popular single-level techniques to multilevel data and introduce MCFA as the natural solution to this problem. We explore the applicability of MCFA for multilevel reliability estimation using simulated data and provide an applied example, focusing on the implications of multilevel reliability estimation for applied researchers. We also provide example code in an appendix in the online supplemental materials to facilitate implementation of the methods we describe.

Single-Level Reliability Estimation Using Structural Equation Modeling

While CFA and structural equation modeling (SEM) estimate construct relations without measurement error when multiple indicators are used for each construct, researchers who use these methods may still wish to estimate reliability for their scales for multiple reasons. First, reliability estimates summarize the factor loading matrix into a single, easily interpretable quantity (given a known number of items) and can help future researchers choose among scales that tap the same construct with better or worse measurement characteristics. Such information may therefore be especially important to provide when discussing the creation and validation of a new scale. Second, providing reliability estimates across many samples can help inform the generalizability of item-level relations (from previous or future research) to their respective latent relations. In the following sections, we begin by discussing how researchers can estimate various reliability coefficients in the framework of CFA and SEM. The list of reliability estimates discussed below is by no means exhaustive, and we readily acknowledge that all estimates carry multiple pros and cons. As such, we constrain our discussion to three of the more commonly utilized reliability estimates: α, ω, and H.

Alpha

Traditional methods of reliability estimation rely on the general linear model (GLM) and are easily implemented in GLM-based frameworks such as SEM and CFA. For example, Cronbach's Equation 16 for computing α (Cronbach, 1951; see also Guttman, 1945; Kuder & Richardson, 1937) specifies α as a function of the average interitem covariance within a scale ($\bar{\sigma}_{ij}$), the variance of the scale score (σ_X^2), and the number of items included in the scale (n):

$$\alpha = \frac{n^2 \bar{\sigma}_{ij}}{\sigma_X^2}. \quad (1)$$

Alpha can be estimated with CFA-capable software by specifying a fully saturated covariance structure model that has no latent variables. The average covariance is found by summing all unique covariances in matrix $\boldsymbol{\Sigma}$ (the symmetric matrix of indicator variances, σ_{ii}^2, and covariances, σ_{ij}) and dividing the sum by the number of unique covariances. The variance of the scale score can then be computed by summing all item variances and two times each unique covariance in $\boldsymbol{\Sigma}$ (the variance of a sum is equal to the sum of the full [i.e., square] covariance matrix of all elements):

$$\sigma_X^2 = \mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}. \quad (2)$$

While it has long been known that α is in most cases an inconsistent estimator of reliability (e.g., Novick & Lewis, 1967), α is by far the most common reliability estimate used in psychological research. The ubiquity of α, as well as the relatively minor difference between α and alternative reliability estimates in applied research (see Footnote 7 below), make alpha an important statistic to consider when examining issues related to scale reliability. We therefore retain α in the present article to make the results of our simulation applicable to a broad audience, for whom the rough approximation provided by α is generally sufficient.

Composite Reliability

The average interitem covariance provides a limited estimate of a scale's true score variance, as evidenced by the fact that α is a consistent estimate of reliability only when all items load on a single underlying construct and when all items represent that construct equally well (i.e., essential tau equivalence; see Novick & Lewis, 1967). CFA allows for heterogeneous correlations between indicators and their underlying common factor(s) (i.e., heterogeneous factor loadings), and *composite reliability* (ω) as calculated from factor loadings produces more precise estimates of reliability than those provided by α.

Composite reliability has been discussed by several authors (e.g., Bentler, 2007; McDonald, 1970, 1999; Raykov, 1997; Werts, Linn, & Jöreskog, 1974) and is conceptually similar to α in that it represents the ratio of a scale's estimated true score variance relative to its total variance. Unlike α, however, ω acknowledges the possibility of heterogeneous item-construct relations and estimates true score variance as a function of item factor loadings (λ_i) in matrix $\boldsymbol{\Lambda}$. Assuming a congeneric scale with a standardized latent construct (i.e., with variance fixed to 1), ω can be estimated as[1]

$$\omega = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \theta_{ii}}, \quad (3)$$

where λ_i represents the factor loading of item i onto a single common factor and θ_{ii} represents the unique variance of item i.

[1] In models with correlated unique factors, the denominator may contain the extra term $2\sum_{i=2}^{k}\sum_{j=1}^{i}\theta_{ij}$, reflecting covariances among unique factors. For simplicity, we focus on the case with uncorrelated unique factors.

The numerator in ω is identical to $\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Lambda}'\mathbf{1}$, the sum of the full (i.e., square) model-implied covariance matrix of item true scores,[2] whereas the denominator represents the true score variance plus all residual variances. Under essential tau equivalence, Equations 1 and 3 become mathematically identical when the factor model used to estimate ω fits the data perfectly. That is, α may be thought of as a special case of ω under essential tau equivalence. The true score covariance matrix contains n^2 elements that all equal the average interitem covariance, and the denominators of both equations simply represent the sum of all sources of scale score variance.

Several variations of composite reliability exist in the literature, most of which reduce to Equation 3 for congeneric scales (e.g., Bentler, 2007; McDonald, 1999; Raykov & Shrout, 2002; Zinbarg et al., 2005). We therefore limit the present discussion to congeneric measures for parsimony.

Maximal Reliability

Composite reliability represents the relation between a scale's underlying latent factor and its unit-weighted composite, but a scale's unit-weighted composite may not optimally reflect its underlying latent construct. The true score variance estimated in factor analysis allows for heterogeneous indicator weights, and it is reasonable to allow similarly heterogeneous weights when creating a scale's composite score. One alternative to comparing true score variance to the variance of a unit-weighted scale is presented as *maximal reliability* (H; e.g., Bentler, 2007; Conger, 1980; Hancock & Mueller, 2001; Li, 1997; Raykov, 2004; see also Thomson, 1940), which represents the reliability of a scale's optimally weighted composite:

$$\mathrm{H} = \frac{\sum_{i=1}^{k} \frac{\ell_i^2}{1-\ell_i^2}}{1 + \sum_{i=1}^{k} \frac{\ell_i^2}{1-\ell_i^2}}, \tag{4}$$

where ℓ_j^2 represents the squared *standardized* factor loading of indicator i onto a single common factor, which is identical to the reliability of indicator i (assuming a correctly specified single-factor model). Hancock and Mueller (2001) showed that this formula reduces to

$$\mathrm{H} = \left(1 + \frac{1}{\sum_{i=1}^{k} \frac{\ell_i^2}{1-\ell_i^2}} \right)^{-1}. \tag{5}$$

Because H optimally weights indicators and squares individual factor loadings, Hancock and Mueller (2001) noted that it has several properties not shared by composite reliability. First, whereas composite reliability is negatively influenced by negative factor loadings (i.e., the numerator sums all factor loadings before squaring the term), the squared loadings used in H allow negatively valenced indicators to contribute meaningful variance to the estimated true score. Second, because H optimally weights indicators when computing the composite score, H will never be less than the reliability (i.e., squared standardized loading) of the best indicator. Similarly, the addition of weakly loading indicators can reduce estimates of composite reliability but will not reduce H because weak indicators will receive very low weights when computing an optimally weighted composite. H therefore considers weak indicators at least *somewhat* informative, and their inclusion in a scale should not reduce the reliability of an optimally weighted composite. This weighting also means, however, that H does not estimate the same population parameter as ω or α.

The coefficients α, ω, and H therefore provide point estimates of a scale's reliability (although H represents the reliability of optimally weighted composite). While point estimates are often informative, it is also important to consider their sampling distributions as our confidence in a point estimate will vary across differing data conditions. The delta-method standard error (i.e., the standard deviation of a parameter estimate's sampling distribution under the assumption of asymptotic normality) has been shown to provide an unbiased estimate of ω's population standard deviation (Raykov, 2002), but the distribution of ω (and other reliability estimates) likely is not symmetric (e.g., Kristner & Muller, 2004, discussed the exact distribution of α and provided an approximation using the *F* distribution). Standard errors may therefore be less informative than empirically derived confidence intervals. For instance, empirically derived confidence intervals would allow researchers to test whether their scale's reliability is significantly greater than a lower threshold such as .80. Raykov (1998) argued that confidence intervals for ω can be obtained through bootstrapping (e.g., Efron & Tibshirani, 1993), and the same logic can also be applied to estimating confidence intervals for α and H.

Multilevel Reliability

A significant body of research has focused on reliability estimation for multilevel models, but research has primarily focused on how reliably group means of a dependent variable represent the larger distribution of group means in a population (e.g., Raudenbush & Bryk, 2002; see also Raykov & Marcoulides, 2006; Raykov & Penev, 2010). While informative for multilevel models in general, the reliability of group means as estimates of the distribution of group means in a population is different than measurement reliability as we have discussed it above.

Of greater present interest is the estimation of a scale's reliability under two-stage random sampling (i.e., multilevel data). Approached from a multilevel perspective, two-stage sampling leads to observed scores (y_{ik}) that contain both true score and measurement error variance at both the within-cluster and between-cluster levels (denoted by i and k subscripts, respectively). Thus, an MCFA approach to multilevel data allows researchers not only to model data for which a scale represents the same construct at each level, but also to model data for which only a between-cluster construct is meaningful (e.g., Chan, 1998; Kozlowski & Klein, 2000) or for which only within-level heterogeneity is meaningful (e.g., Fitzmaurice, Laird, & Ware, 2011; Halaby, 2004; Woolridge, 2002). Further, MCFA allows for qualitatively different constructs at each level such that a single scale may contain items that possess different factor structures within versus between clusters.

Whereas the concept of separate level-specific true scores and measurement errors at each level runs contrary to the tenets of

[2] This formula is slightly different for multidimensional scales. Zinbarg, Revelle, Yovel, and Li (2005) distinguished between composite reliability for congeneric versus hierarchical scales, for instance.

classical test theory (e.g., that there is only one measurement error), we note that multilevel models represent a superordinate class of models that include those derived from generalizability theory (see Goldstein & McDonald, 1988, for a brief discussion). While issues concerning reliability are more appropriately discussed in terms of classical test theory, generalizability theory specifies that a scale's total variance can be decomposed into multiple orthogonal facets (e.g., Shavelson & Webb, 2004; Webb, Rowley, & Shavelson, 1988), similar to the decomposition that occurs under the MCFA model. A generalizability theory-derived model may, for instance, decompose a scale's variance into the variance of person-specific deviations from a grand mean (i.e., a universe score), item-specific variance, between-cluster differences, the interactions among these three sources, as well as variance due to nonsystematic error. Here, nonsystematic variation and variance due to the three-way interaction between persons, items, and clusters are not distinguishable, so these effects would be aggregated into a single residual term. Using such a decomposition, researchers can calculate generalizability coefficients that treat different combinations of these sources of variance as representing target variance versus error. Generalizability coefficients are therefore akin to the reliability estimates discussed in classical test theory.

The MCFA model assumes a similar kind of decomposition. The MCFA model decomposes observed information in an item into components related to each individual's cluster-average true score (T_{bk}), which characterizes all individuals within cluster *k,* as well as each individual's true deviation from the cluster-average true score (T_{wi}). Furthermore, the difference between an individual's within-cluster deviation from the cluster average and that individual's T_{wi} is a within-cluster error ($E_{wi} = y_{ik} - y_{.k} - T_{wi}$). The difference between a cluster's deviation from the grand mean of all true scores and that cluster's T_{bk} is a between-cluster error ($E_{bk} = y_{.k} - T_{bk}$). Thus, we may represent an individual's observed score as the sum of four parts:

$$y_{ik} = \underbrace{T_{wi} + E_{wi}}_{\text{within-cluster}} + \underbrace{T_{bk} + E_{bk}}_{\text{between-cluster}} . \quad (6)$$

Table 1 clarifies how each of these elements aligns with the variance components that would exist in a parallel model derived from generalizability theory. The assumption of distinct within- and between-cluster true scores suggests that true score variance can be captured to a different degree at each level, motivating the need for separate reliability estimates at each level. Reliability at the within level represents the ratio of the within-cluster true score variance to total within-cluster variance ($\text{var}(T_{wi})/\text{var}(T_{wi} + E_{wi})$), whereas reliability at the between level represents the ratio of the between-cluster true score variance to total between-cluster variance ($\text{var}(T_{bk})/\text{var}(T_{bk} + E_{bk})$). Between-cluster reliability therefore reflects the reliability of the between-cluster information in a scale and does not necessarily represent the reliability of group-level composites.

Between-cluster reliability is therefore distinct from an intraclass correlation (ICC), which represents the ratio of a scale score's between-cluster variance relative to its total variability across both levels. Between-cluster reliability instead reflects the degree to which group-level differences in a researcher's observed data can be generalized to represent between-group differences in a construct of interest.

Table 1
Relations Between Generalizability Theory Facets and Parallel Elements of the Multilevel Confirmatory Factor Analysis Model

Facet	Element of multilevel confirmatory factor analysis model
Person	Within-cluster true score
Cluster	Between-cluster true score
Item	No variance across observations; represented in item intercepts
Person × Cluster	Undefined under the assumption of no cross-classification
Person × Item	Within-cluster error; also includes nonsystematic within-cluster variance (error)
Cluster × Item	Between-cluster error; also includes nonsystematic between-cluster variance (error)
Person × Cluster × Item	Undefined under the assumption of no cross-classification

To the extent that group-level reliability estimates are applicable across studies, estimates of between-group reliability in one study help inform the validity of group-level inferences based on previous or future studies. It may therefore be especially important to estimate multilevel reliability estimates when constructing new scales to show that a new scale reliably captures true score variation at each possible level of analysis. Providing such estimates will therefore allow researchers to make better informed choices between measurement instruments, especially in the context of multilevel hypotheses.

An MCFA approach to estimating level-specific reliability is especially important to researchers dedicated, for one reason or another, to the multilevel analysis of scale composites. For instance, a multilevel SEM might not be estimable with small unit-level sample sizes, but multilevel regression with scale scores would be. Although not as ideal as fitting a full multilevel SEM (MSEM) model to all scales in a battery simultaneously, separate MCFA models for each scale in such a study could be used to estimate multilevel reliability for each scale individually. This MCFA approach would be greatly preferable to the currently available options of either (a) not reporting reliability and hoping for the best or (b) assuming a single-level design and computing a single-level reliability estimate. Ignoring the fact that separate reliabilities may exist at each level of analysis by computing single-level reliability conflates within- and between-cluster reliability, and it can be shown that single-level reliability is a simple mathematical function of a scale's ICC and its reliability at each level of analysis (see Appendix A in the online supplemental materials).

Raykov and du Toit (2005) provided one means for estimating composite reliability in multilevel CFA that accounts for variability both within and between groups, but their procedure provided only a single estimate of reliability that does not differentiate reliability within groups from reliability between groups. A single estimate of composite reliability provides information about the overall reliability of a scale but does not inform whether the scale is sufficiently reliable for use at a specific level of analysis.

Cranford and colleagues (2006) also addressed the issue of multilevel reliability, suggesting a method that indeed produces separate level-specific reliability estimates. Their method conflates within- and between-cluster error variance, however, and is not as generalizable as a method that acknowledges separate true score and error variances at each level.

Instead, MCFA allows separate estimation of level-specific measurement model parameters and thus allows for level-specific reliability estimation. The general CFA framework allows estimation of α, ω, and H, as well as other reliability estimates we do not discuss. We next describe how each can be extended to a two-level context.[3]

A Multilevel CFA Approach

The extension of CFA to accommodate two-level data allows separate estimation and analysis of within- and between-cluster covariance matrices. As discussed above, CFA is optimal for estimating reliability in single-level data, and an MCFA extension to reliability estimation is relatively straightforward (see also Raykov & du Toit, 2005).[4] In the remainder of this article, we discuss the estimation of α, ω, and H in an MCFA framework.

There are several approaches in the methodological literature for conducting MCFA (e.g., Muthén, 1990, 1994). Here, we adopt a method recently developed by Muthén and Asparouhov (2009, 2011) for conducting MSEM. MCFA is a special case of MSEM with no structural paths linking latent variables, in much the same way that single-level CFA is a special case of SEM.

Briefly, the MCFA model is given as a special case of Muthén and Asparouhov's (2009) model by a set of three equations (retaining their notation):

$$\mathbf{Y}_{ik} = \mathbf{\Lambda}_k \boldsymbol{\eta}_{ik}, \tag{7}$$

$$\boldsymbol{\eta}_{ik} = \boldsymbol{\alpha}_k + \mathrm{B}_k \boldsymbol{\eta}_{ik} + \boldsymbol{\zeta}_{ik}, \tag{8}$$

$$\boldsymbol{\eta}_k = \boldsymbol{\mu} + \boldsymbol{\beta} \boldsymbol{\eta}_{ik} + \boldsymbol{\zeta}_k, \tag{9}$$

where i and k index cases (Level 1 units) and clusters (Level 2 units), respectively. $\mathbf{Y}_{ik}$ is a vector of p measured variables; $\mathbf{\Lambda}_k = \mathbf{\Lambda} = [\mathbf{I}_p\ \mathbf{0}_{p\times m}\ \mathbf{I}_p\ \mathbf{0}_{p\times m}]$ is a $(p \times (2p + 2m))$ factor loading matrix linking $\mathbf{Y}_{ik}$ to p latent parts at both the within- and between-cluster levels and m common factors at both levels; $\boldsymbol{\eta}_{ik}$ is a vector of length $(2p + 2m)$ containing p latent within-cluster parts, m within-cluster common factors, p latent between-cluster parts, and m between-cluster common factors[5]; $\boldsymbol{\alpha}_k$ is a vector of length $(2p + 2m)$ that contains the p item intercepts and m between-cluster common factors; $\mathbf{B}_k$ is a $(2p + 2m) \times (2p + 2m)$ matrix containing within-cluster factor loadings; $\boldsymbol{\eta}_k$ $(r \times 1)$ contains all of the k-subscripted random coefficients from $\boldsymbol{\alpha}_k$ and $\mathbf{B}_k$, including the between-cluster common factors; $\boldsymbol{\mu}$ $(r \times 1)$ contains means of those coefficients and the item intercepts (if desired); $\boldsymbol{\beta}$ $(r \times r)$ contains between-cluster factor loadings; $\boldsymbol{\zeta}_{ik}$ contains unique factors and common factor residuals for the within-cluster model; and $\boldsymbol{\zeta}_k$ $(r \times 1)$ contains unique factors and common factor residuals for the between-cluster model. Finally, $\boldsymbol{\zeta}_{ik} \sim MVN(\mathbf{0}, \mathbf{\Psi}_W)$, and $\boldsymbol{\zeta}_k \sim MVN(\mathbf{0}, \mathbf{\Psi}_B)$. A fully specified path diagram is included in Appendix B in the online supplemental materials (for the case in which the item intercepts $\tau_j = 0$).

Whereas the basic MCFA model can be elaborated in various ways, we restrict our focus to factor models with no covariates, only continuous items, and no latent regressions (apart from paths among latent variables and their indicators that are better conceptualized as loadings than as regression weights). Furthermore, we consider only the case in which item intercepts are omitted, factor loadings do not vary randomly at the cluster level ($\mathbf{B}_k = \mathbf{B}$), and the configural factor structure is identical across levels. These simplifications yield constrained versions of Equations 7 and 8:

$$\mathbf{Y}_{ik} = \mathbf{\Lambda} \boldsymbol{\eta}_{ik}, \tag{10}$$

$$\boldsymbol{\eta}_{ik} = \boldsymbol{\alpha}_k + \mathbf{B} \boldsymbol{\eta}_{ik} + \boldsymbol{\zeta}_{ik}, \tag{11}$$

Appendix B in the online supplemental materials presents an example model with full expansion of all matrices.

Multilevel alpha. As discussed above, α, ω, and H can be directly estimated from CFA model parameters, suggesting a simple extension to two-level CFA. Separate within- and between-cluster α can be obtained by specifying fully saturated indicator covariance matrices in both levels of an MCFA and separately applying Equation 1 to the within- and between-cluster results. The numerator of each level-specific α is therefore the squared number of indicators present at a given level of analysis multiplied by the average covariance at that same level. The denominator of each level-specific α similarly represents the sum of all elements in the full (i.e., square) level-specific covariance matrix and can be obtained by summing all level-specific indicator variances and two times each unique level-specific covariance.

Multilevel composite reliability. Extending ω to MCFA requires specification of a unidimensional factor structure at both the within- and between-cluster levels. Equation 3 is then applied to the level-specific parameter estimates, making within-level ω a function of within-level factor loadings and residual variances whereas between-cluster ω is a function of the between-cluster factor loadings and residual variances. This approach therefore requires that residual variances be estimated at both levels instead of fixing Level 2 residual variances to zero, as is sometimes done (e.g., Gottfredson, Panter, Daye, Allen, & Wightman, 2009). Fixing Level 2 residual variances to zero assumes perfect Level 2 reliability, rendering Level 2 reliability estimation unnecessary.

While beyond the scope of the present article, we note that unless a researcher has strong reasons to suspect an item is perfectly reliable at the between level, he or she should avoid fixing

[3] See also Wilhelm and Schoebi (2007), who presented a similar method of level-specific reliability estimation in an multilevel modeling framework. Our method differs from theirs in that we estimate true score and error variances at each level using MCFA, allowing for estimates of level-specific α, ω, and H. Wilhelm and Schoebi's approach specifies level-specific reliability as a function of level-specific covariances and therefore only allows for the estimation of level-specific α.

[4] Our approach specifically considers cases for which a researcher wishes to know reliability at two levels. When reliability at the between level is not relevant to a researcher, alternative approaches such as estimating reliability using group-mean-centered data may represent a simpler alternative. We thank an anonymous reviewer for suggesting this possibility. Huang and Weng (2012) presented a similar approach to estimating reliability in ecological momentary assessment data.

[5] Initially, it may appear strange to see $\boldsymbol{\eta}_{ik}$ on both sides of this equation. This simply denotes that some elements of $\boldsymbol{\eta}_{ik}$ are functions of other elements of $\boldsymbol{\eta}_{ik}$.

Level 2 residual variances to zero. Estimated residual variances may be close to zero when an item's ICC is small, but near-zero residual variances can still represent a substantial proportion of an indicator's between-cluster variability (e.g., 10%) if the corresponding Level 2 true score variance is also near zero. For example, the population model discussed by Hox, van de Schoot, and Matthijsse (2012; see also Meuleman & Billiet, 2009) specifies between-cluster factor loadings of .265 and .280 for a between-cluster latent construct that has been standardized to have unit latent variance. The respective item ICCs are relatively low (i.e., .080) such that the residual variances of these two indicators are close to zero (.015 and .008, respectively). Despite being very small, these residual variances correspond to approximately 19% and 10% of the total between-cluster item variances. In this case, assuming perfect between-cluster item reliability (i.e., fixing residuals to zero) would not be justifiable.

Multilevel maximal reliability. The same model used to estimate level-specific ω can also be used to estimate level-specific H. H requires estimates of each item's reliability, as provided by squared standardized factor loadings (represented in Equations 4 and 5 as ℓ_i^2). Standardized factor loadings are not generally provided by MCFA software, and a simple generalization allows H to be estimated from raw-metric parameter estimates:

$$\frac{\ell_i^2}{1-\ell_i^2} = \frac{\lambda_i^2}{\sigma_i^2 - \lambda_i^2} \approx \frac{\lambda_i^2}{\theta_{ii}}. \quad (12)$$

Computation of level-specific reliability is straightforward from an MCFA perspective, but the applicability of this method remains untested. Next, we present a simulation that explores the use of multilevel α, ω, and H and shows how these measures capture level-specific reliability under several conditions. We additionally calculate single-level estimates of α, ω, and H to show the potential implications of ignoring nested data structures when computing reliability.

Confidence intervals and credible intervals. Following Raykov's (1998) logic, we note that confidence intervals for level-specific reliability should be obtainable via bootstrapping or similar procedures. Direct nonparametric bootstrapping (i.e., iteratively resampling from an empirical data set to derive an empirically based estimate of a parameter's sampling distribution) can produce biased estimates when applied to multilevel data, however, and parametric alternatives are preferable when one is willing to accept all model assumptions (Goldstein, 2011). We therefore examine two approaches to estimating the sampling distribution of our level-specific reliability estimates: Monte Carlo confidence intervals and Bayesian credible intervals. As we explain below, Monte Carlo confidence intervals are derived by generating random draws from a parameter vector's hypothesized joint distribution and computing functions of these parameters on each draw. Bayesian credible intervals are similarly derived from a parameter vector's posterior distribution rather than being computed directly from resampled data.

Monte Carlo confidence intervals. Monte Carlo confidence limits are derived by relying on the sampling distribution of parameter estimates (e.g., factor loadings and residual variances), generating random draws for each parameter, and computing a statistic of interest (e.g., a reliability estimate) from the generated parameters. This is different from a parametric bootstrap approach, where data sets are generated from a model and statistics are computed for each data set separately. An elaboration of the differences between Monte Carlo and parametric bootstrap methods is beyond the scope of the present article; we refer readers to Preacher and Selig (2012) for a discussion of this topic.

We obtained Monte Carlo confidence intervals by drawing a random sample of 10,000 parameter estimates from an estimated sampling distribution of these estimates, where observed parameter estimates were used as distribution means and the asymptotic covariance matrix of these estimates was used to represent their population covariance matrix. While Monte Carlo confidence intervals may be accurate with substantially fewer than 10,000 random draws (e.g., MacKinnon, Lockwood, & Williams, 2004, suggested 1,000 draws), the computational intensity of this approach is small. Because of this, we chose 10,000 draws as an arbitrarily large number that will provide a precise estimate of our target parameters' expected sampling distributions while costing little in terms of computational power and time.

Given these population parameters and a set of distributional assumptions (i.e., that all parameters were multivariate normally distributed), we drew observations from the joint distribution of parameter estimates and computed reliability estimates using each sample of parameter estimates. We provide R syntax for computing Monte Carlo confidence intervals for level-specific reliability estimates, given a set of MCFA factor loadings and residual variances and their corresponding asymptotic covariance matrix. Our syntax can be found in Appendix C in the online supplemental materials.

Bayesian credible intervals. Bayesian credible intervals are the range of parameter values that best describe a set of data with a fixed level of probability, given a set of priors. Thus a 95% credible interval for parameter θ indicates the range of values that θ has a 95% probability of falling between, given observed data and priors. While this definition differs from the frequentist concept of a confidence interval, credible intervals with noninformative priors approximate asymmetric nonparametric bootstrap confidence intervals (see DiCiccio & Efron, 1996, p. 211). Raykov (1998, e.g.) has shown that nonparametric bootstrapping can produce unbiased estimates of a reliability estimate's sampling distribution, but resampling-based confidence intervals are not appropriate for multilevel data (Goldstein, 2011), and bootstrapping generally is problematic in cases involving computational difficulties where a Bayesian approach with Markov chain Monte Carlo can work well (Efron, 2011). Because credible intervals are computed as a function of posterior distributions rather than from direct resampling, we expect close agreement between Bayesian credible intervals (calculated using the ESTIMATOR = BAYES option in Mplus) and the corresponding quantiles of a parameter's sampling distribution.

Hypotheses

Given the above discussion, we can make hypotheses regarding the performance of single- and multilevel reliability estimates when data adhere to a multilevel structure. First, ignoring nested data structures will render single-level reliability estimates difficult to interpret when reliability is not identical across levels. Single-level reliability necessarily averages across levels of measurement (see also Appendix A in the online supplemental materials), so we anticipate that single-level reliability estimates will more closely reflect within-level reliability as the proportion of the

scale's variance shifts toward the within level (i.e., as ICC becomes smaller) and will more closely reflect between-cluster reliability as the proportion of the scale's variance shifts toward the between level (i.e., as ICC becomes larger; Hypothesis 1A). When reliability does not differ across levels, however, single-level reliability should simply represent the weighted average of two identical values. We hypothesize that when the level-specific reliabilities do not differ across levels, single-level reliability estimates will be unbiased with respect to actual level-specific reliability at either level (Hypothesis 1B).

Second, we can make specific hypotheses about the relative performance of individual reliability estimates, including bias, model convergence rates, and confidence interval coverage. As for performance of the individual measures, factor analysis can fail to reproduce an underlying factor structure when item reliabilities are low, especially when sample size is also low (e.g., MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum, Widaman, Zhang, & Hong, 1999). We will calculate reliability estimates using either a saturated model (α) or by fitting the data-generating model (ω and H) and anticipate generally low levels of bias for two-level reliability estimates. In line with previous research (e.g., Muthén, Kaplan, & Hollis, 1987), we consider absolute percent bias < 10 to represent an acceptable level of bias. We do, however, anticipate increased bias when level-specific sample size or reliability is low (Hypothesis 2).

Third, ω differs from α only by relaxing the assumption of essential tau equivalence. Estimates of ω and α consequently are very similar under moderate violations of essential tau equivalence,[6] and despite the widely acknowledged limitations of α (e.g., Sijtsma, 2009), we anticipate that the two estimates will perform similarly (Hypothesis 3A). H differs from both α and ω by representing the reliability of an optimally weighted scale score rather than the reliability of a unit-weighted composite. We therefore anticipate that H will be influenced by different factors than α and ω (Hypothesis 3B).

Fourth, we will consider convergence rates for each model. We estimate our level-specific αs using a fully saturated two-level model and anticipate few if any convergence problems (Hypothesis 4A). Unlike α, ω and H are calculated by specifying a single-factor model at both levels, and model convergence is not guaranteed. We anticipate that the two-level CFA models needed to estimate two-level ω and H will converge in the overwhelming majority of trials when sample size is sufficiently large at both levels. Common rules of thumb suggest that SEM models generally require more than 100 or 150 observations (e.g., Brown, 2006), which we apply as a sufficiently large sample size at the within level. Meuleman and Billiet (2009) suggested that 40 between-cluster observations may be sufficient for simple models but that 60 or even 100 groups may be required to achieve sufficient power to detect small interconstruct relations at the between-cluster level. We anticipate reduced convergence rates when overall sample size is small, especially when the between-cluster sample size is small (i.e., 50) or the overall sample size is small due to having dyadic data and few (50 or 100) between-cluster groups (Hypothesis 4B), or when item reliabilities are low at either level (Hypothesis 4C).

Our final hypotheses concern the Monte Carlo confidence limits and Bayesian credible intervals discussed above. Under the assumption of multivariate normality (which our simulations also assume), we anticipate that Monte Carlo confidence limits generated by our R code (see Appendix C in the online supplemental materials) will closely correspond to empirically derived confidence limits based on the sampling distribution of level-specific reliability estimates. In other words, we anticipate that 95% confidence limits generated using our calculator should provide unbiased estimates of the 2.5th and 97.5th percentiles of the reliability estimates' actual sampling distributions (Hypothesis 5). Further, we also anticipate that 95% credible intervals estimated using Bayesian analysis will similarly provide unbiased estimates of the actual 2.5th and 97.5th percentiles of the reliability estimates' sampling distributions (Hypothesis 6).

Simulation

Our simulation considers calculations of single-level and multilevel α, ω, and H under conditions when data originate from a known multilevel structure. We examine reliability for a congeneric six-item scale with conditions varying the number of Level 1 units, the number of Level 2 units, the average item ICC, and the level(s) of analysis at which the indicators show high reliability (henceforth, the reliability condition; see Table 2). We generated 1,000 replications for each condition and analyzed our data using robust maximum likelihood estimation in Mplus. We chose 1,000 replications as an arbitrary balance between generating a large enough number of replications to obtain appropriate precision of our estimates and the time required to analyze our models. We also chose 1,000 replications per cell because this number has been used in previous simulation studies (e.g., Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009).

For each condition, we separately calculated the percent bias[7] of single-level α, ω, and H relative to the actual level of each at the within and between levels; percent bias for the estimates of within- and between-cluster α, ω, and H; and convergence rates for our two-level models. We then conducted analyses of variance (ANOVAs) to determine which conditions most strongly influenced each estimate and convergence rate. We implemented a tear-down approach to decide which predictors were retained in the final models, retaining predictors and interactions only if they explained at least 5% of the dependent variable's variance (i.e., removing the predictor or interaction caused a decrease in η^2 greater than or equal to .05). We chose a 5% change in η^2 as an arbitrary value that balances the need to detect small but important interactions with our desire to avoid interpreting statistically significant interactions that would nevertheless have little bearing on applied applications of the above reliability estimates. In line with previous research (e.g., Muthén et al., 1987), we also consider absolute percent bias $< 10\%$ to represent an acceptable level of bias.

To test our expectation that Monte Carlo confidence intervals will closely reflect the distribution of reliability estimates across repeated sampling, we also obtained 95% confidence limits for several conditions using the R code provided in Appendix C in the online supple-

[6] For example, a construct with six indicators having $\lambda_1 = \lambda_2 = .80$, $\lambda_3 = \lambda_4 = .60$, and $\lambda_5 = \lambda_6 = .40$ obviously violates the assumption of essential tau equivalence. Despite this, the difference between α and ω for this construct is trivial (α = .77, ω = .78).

[7] Because reliability estimates are not anticipated to be normally distributed, the median parameter estimate in each condition better represents the statistic's central tendency than the mean parameter estimate. Bias in each condition was therefore computed as ([median estimate − parameter]/parameter) × 100.

Table 2
Simulation Conditions

Observations per cluster: 2, 15, 30
Number of clusters: 50, 100, 200
Indicator intraclass correlation: .05, .25, .50, .75
Level(s) of analysis with high reliability (reliability condition):
Within only, between only, both levels, neither level
Level(s) with high reliability: $\lambda_1 = \lambda_2 = .80$; $\lambda_3 = \lambda_4 = .70$; $\lambda_5 = \lambda_6 = .60$; $\alpha = .852$, $\omega = .854$, $H = .868$
Level(s) without high reliability: All $\lambda_j = .30$; $\alpha = \omega = H = .372$

mental materials. We then compared these limits to the empirical distribution of reliability estimates obtained from our simulation. These conditions involved either a low or high ICC (i.e., .05 vs. .50, respectively), varied the total number of observations (either 200 clusters with 30 observations each or 100 clusters with two observations each), and examined conditions where reliability was high at

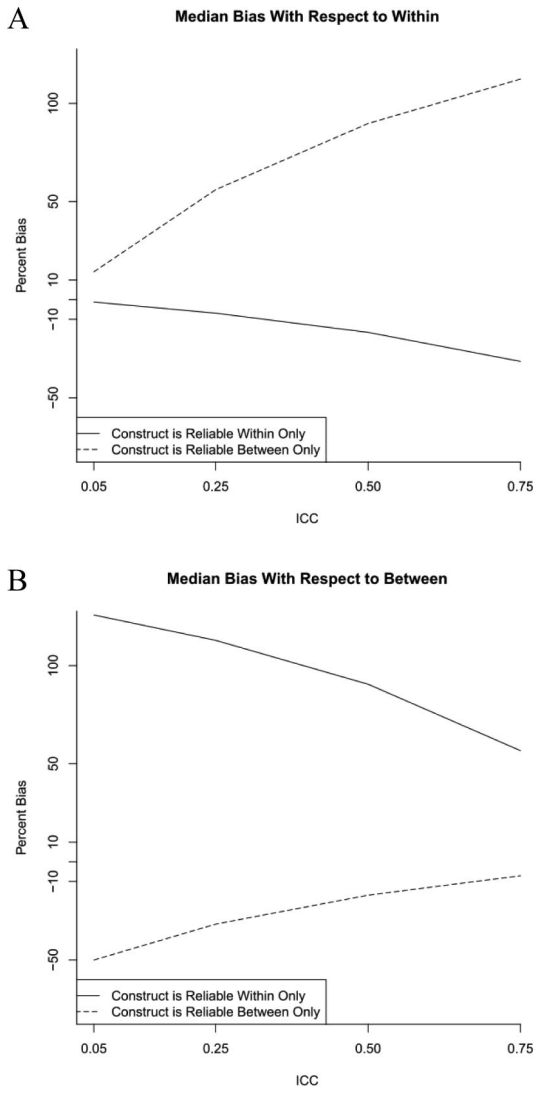

Figure 1. Bias of single-level α with respect to actual reliability within and between. ICC = intraclass correlation.

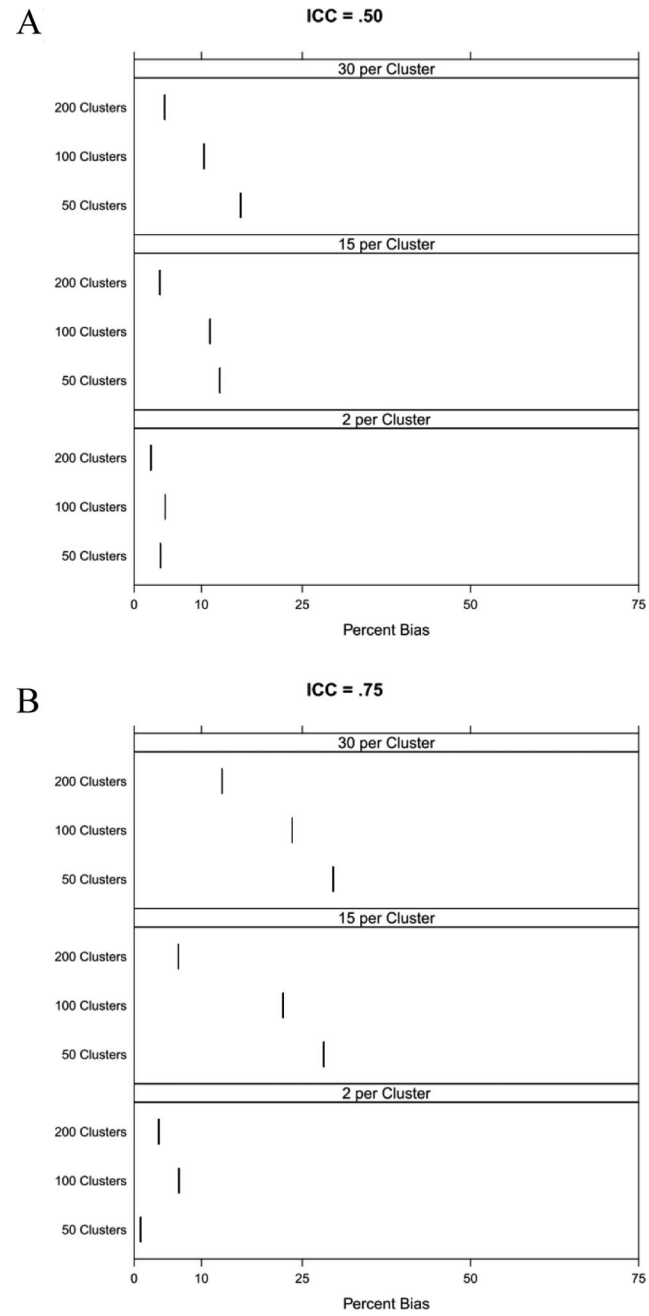

Figure 2. Bias of single-level ω when the scale was not reliable at either level. ICC = intraclass correlation.

both levels, neither level, only within, or only between. We selected these specific cases to present readers with a general sense of how Monte Carlo confidence intervals behave in relatively extreme cases (i.e., high vs. low ICCs, large vs. small samples) without providing an unwieldy amount of information. We then test Hypothesis 6, that Bayesian credible intervals will similarly produce unbiased estimates of the reliability estimates' sampling distributions, by fitting Bayesian models in Mplus to data generated to adhere exactly to the population parameters.

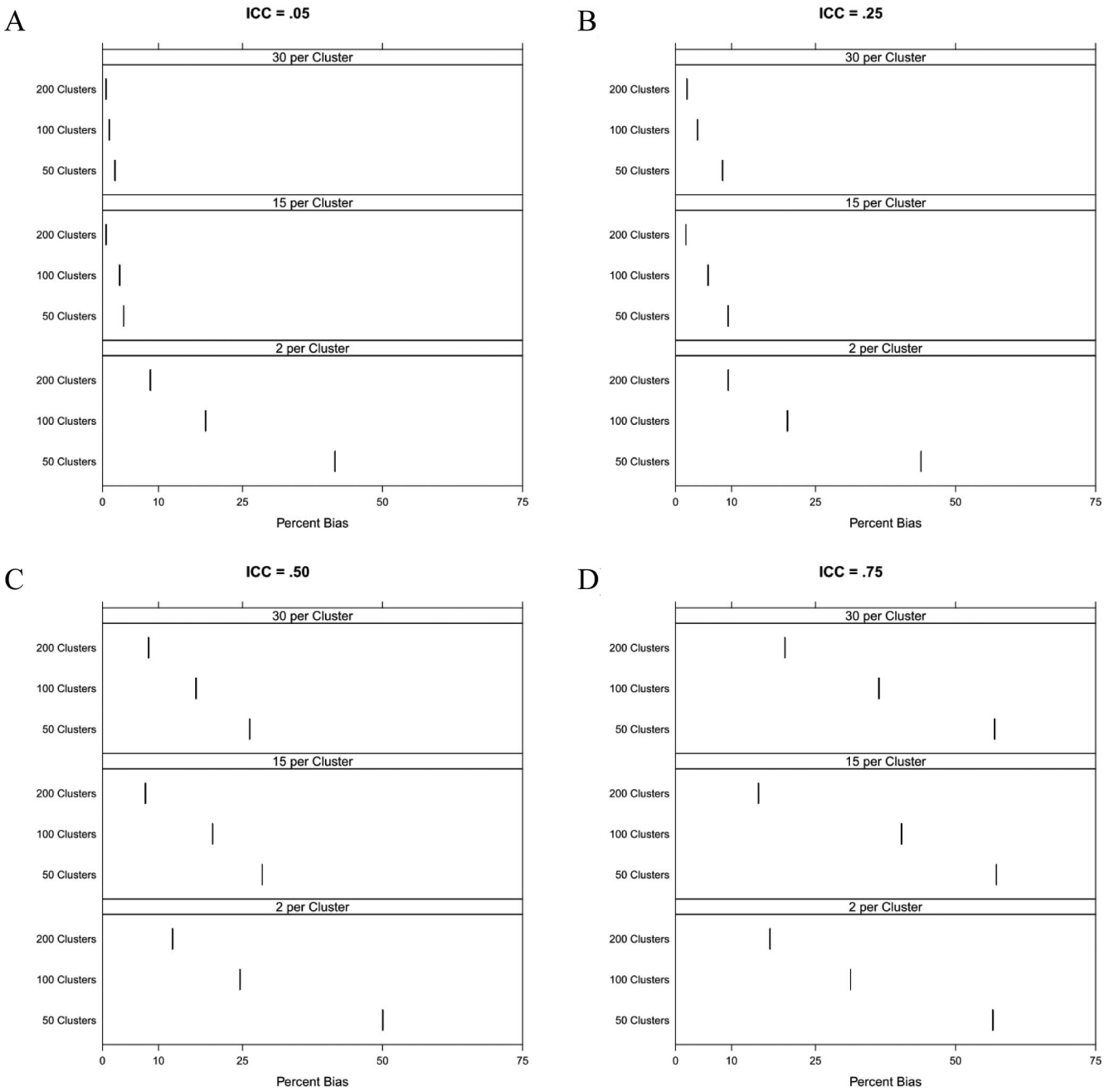

Figure 3. Bias of single-level H when the scale was not reliable at either level. ICC = intraclass correlation.

Results

We present results from our simulation below. All results consider data only from models that converged and had acceptable fit (i.e., root-mean-square error of approximation [RMSEA] $< .08$, comparative fit index [CFI] $> .90$, and Tucker-Lewis index [TLI] $> .90$); these results did not meaningfully differ from results that also included models that converged but did not display acceptable fit. As discussed below, convergence rates were generally high but suffered in conditions with very weak data (e.g., low reliability at the between level, small cluster sizes).

Single-Level Results

All three single-level reliability estimates were significantly predicted by a two-way interaction between ICC and the reliability condition, α: $F(15, 143595) = 5.30 \times 10^7$, $p < .001$, $\eta^2 = .999$; ω: $F(15, 109649) = 1.41 \times 10^7$, $p < .001$, $\eta^2 = .999$; H: $F(15, 109649) = 170522$, $p < .001$, $\eta^2 = .959$. Results for all three reliability estimates supported Hypothesis 1A, with single-level estimates generally biased as a function of ICC when the scale was reliable at only one of the two levels. With

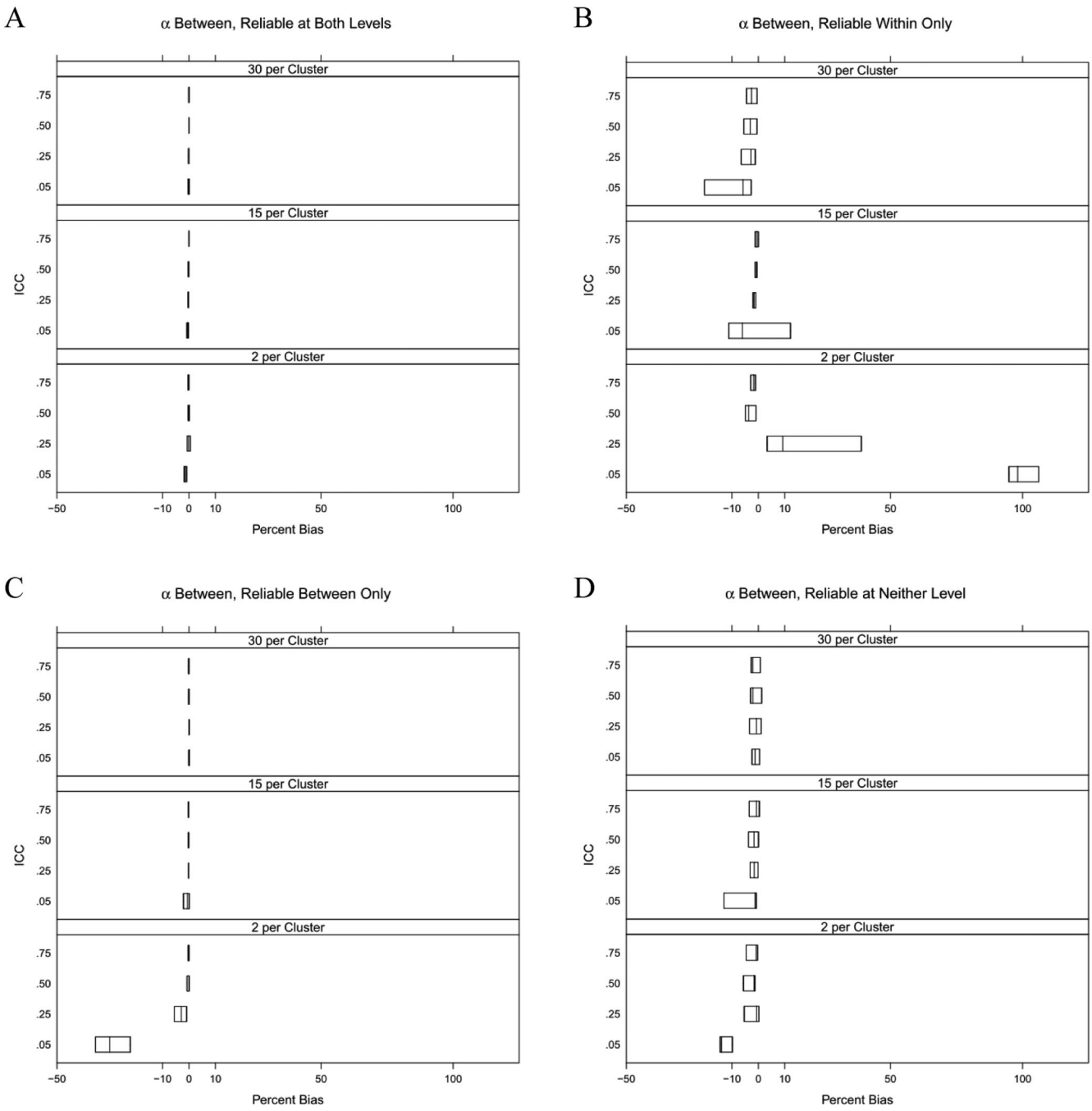

Figure 4. Bias of between-cluster α. ICC = intraclass correlation.

respect to actual reliability at the within level, single-level estimates were negatively biased when the average item ICC increased above .25 and the construct was reliable at only the within level. When the construct was reliable at only the between level, single-level estimates were always positively biased relative to the actual within-level reliability.

We found similar, but inverse, relations when reliability estimates were considered in respect to actual reliability at the between level. Single-level estimates were positively biased when the construct was reliable at only the within level and negatively biased when the construct was reliable at only the between level, with the bias becoming less severe as ICC increased. Figure 1 graphically displays the results for single-level α and is representative of the results for both ω and H.

Our results also generally supported Hypothesis 1B, which predicted that single-level reliability estimates would consistently estimate reliability at both levels when the actual reliability at both levels was in fact the same. Estimates of α never displayed

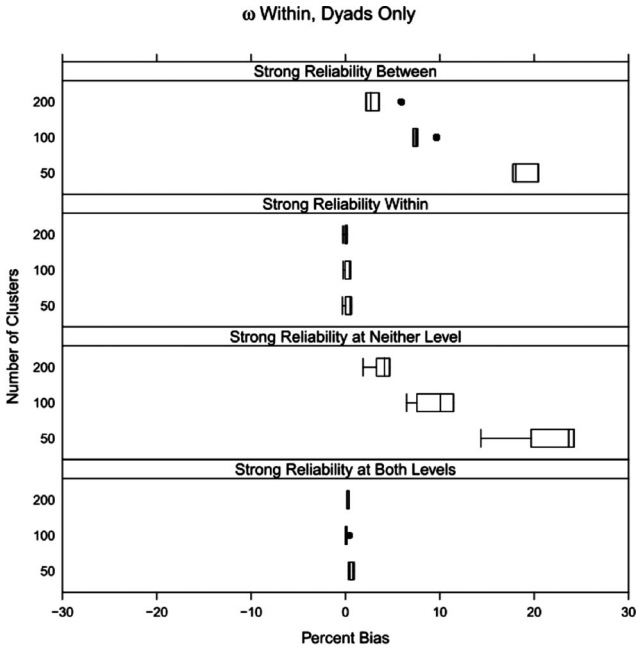

Figure 5. Bias of within-level ω.

absolute percent bias greater than 10% when the scale was equally reliable at both levels, nor did either ω or H when the scale displayed high reliability at both levels. When reliability was low at both levels, both ω and H displayed a strong upward bias in many instances, however. Follow-up ANOVAs found that when the scale was not reliable at either level, the bias in ω was predicted by a three-way interaction of ICC, the number of clusters, and the number of observations per cluster. This three-way interaction accounted for all possible between-cell differences, and our F statistic was undefined with an η^2 of 1.00. Estimates of ω displayed substantial positive bias only when ICC was .50 or higher and, as Figure 2 shows, increased as ICC increased, the number of observations per cluster increased, and the number of clusters decreased.

Similar ANOVAs found that when the scale was not reliable at either level, bias in H was predicted by a main effect of ICC as well as a two-way interaction between the number of clusters and the number of observations per cluster, $F(11, 18075) = 63{,}669.70$, $p < .001$, $\eta^2 = .975$. Figure 3 shows that H became increasingly biased as ICC increased and as both the number of clusters and the number of observations per cluster decreased.

Two-Level Alpha

Within-level α never displayed percent bias greater than 10% and was considered to be acceptable in all conditions. Between-cluster α was biased in several conditions, however (see Figure 4). Bias in between-level α was predicted by a three-way interaction of ICC, the reliability condition, and the number of observations per cluster, $F(47, 143431) = 63{,}954.00$, $p < .001$, $\eta^2 = .954$. In general, between-cluster α was negatively biased for small clusters when ICC was low and reliability at the within level was low. Between-cluster α was also positively biased for small clusters when ICC was low and when reliability at the within level only was high. Combined, these results suggest that between-cluster α is pulled toward the within-level reliability when the between-cluster covariance matrix is informed by a limited number of observations per cluster and there is low between-cluster variability, partially supporting our Hypothesis 2.

Two-Level Composite Reliability

Bias in within-level ω was predicted by a three-way interaction of the reliability condition, the number of clusters, and the number of observations per cluster, $F(35, 125155) = 112{,}933.00$, $p < .001$, $\eta^2 = .969$. Follow-up analyses revealed that within-level ω displayed unacceptable bias only in conditions with two observations per cluster and low reliability at the within level. As Figure 5 shows, bias increased under these conditions as the number of clusters decreased. This finding again supports Hypothesis 2: Within-level ω appeared to be positively biased when overall sample size was low (indicated here by a decreasing number of dyads) and when reliability at the within level was also low.

Bias in between-cluster ω was predicted by a three-way interaction of ICC, the reliability condition, and the number of observations per cluster, $F(47, 125143) = 68{,}550.50$, $p < .001$, $\eta^2 = .963$. As Figure 6 shows, between-cluster ω was positively biased when the between-cluster reliability and the item ICCs were both low, especially when the number of observations per cluster was small. As with within-level ω, this finding supports our Hypothesis 2: Between-cluster ω was positively biased when there were few observations per cluster (i.e., small samples), when between-cluster reliability was low, and when the Level 2 variances were small (i.e., low ICCs).

Two-Level Maximal Reliability

Bias in within-level H was predicted by the three-way interaction of the reliability condition, the number of clusters, and the number of observations per cluster, $F(35, 125155) = 356{,}143$, $p < .001$, $\eta^2 = .990$. As with within-level ω, results indicated that within-level H was positively biased for dyadic data with a low overall sample size (i.e., dyads) and when the within-level reliability was low (see Figure 7). These results again support our Hypothesis 2.

Bias in between-cluster H was significantly predicted by two-way interactions between the reliability condition and both ICC and the number of clusters, $F(23, 125167) = 37{,}842.7$, $p < .001$, $\eta^2 = .874$. As Figure 8 shows, between-cluster H was extremely positively biased when the between-cluster reliability was low, especially when item ICCs were also low and there were few observations per cluster. This extreme positive bias when between-cluster H is actually low in the population suggests that the sample estimate of H is not a consistent estimator of its population value.

Relative Performance

Our Hypothesis 3 predicted that α and ω would perform similarly (Hypothesis 3A) but that different factors would pre-

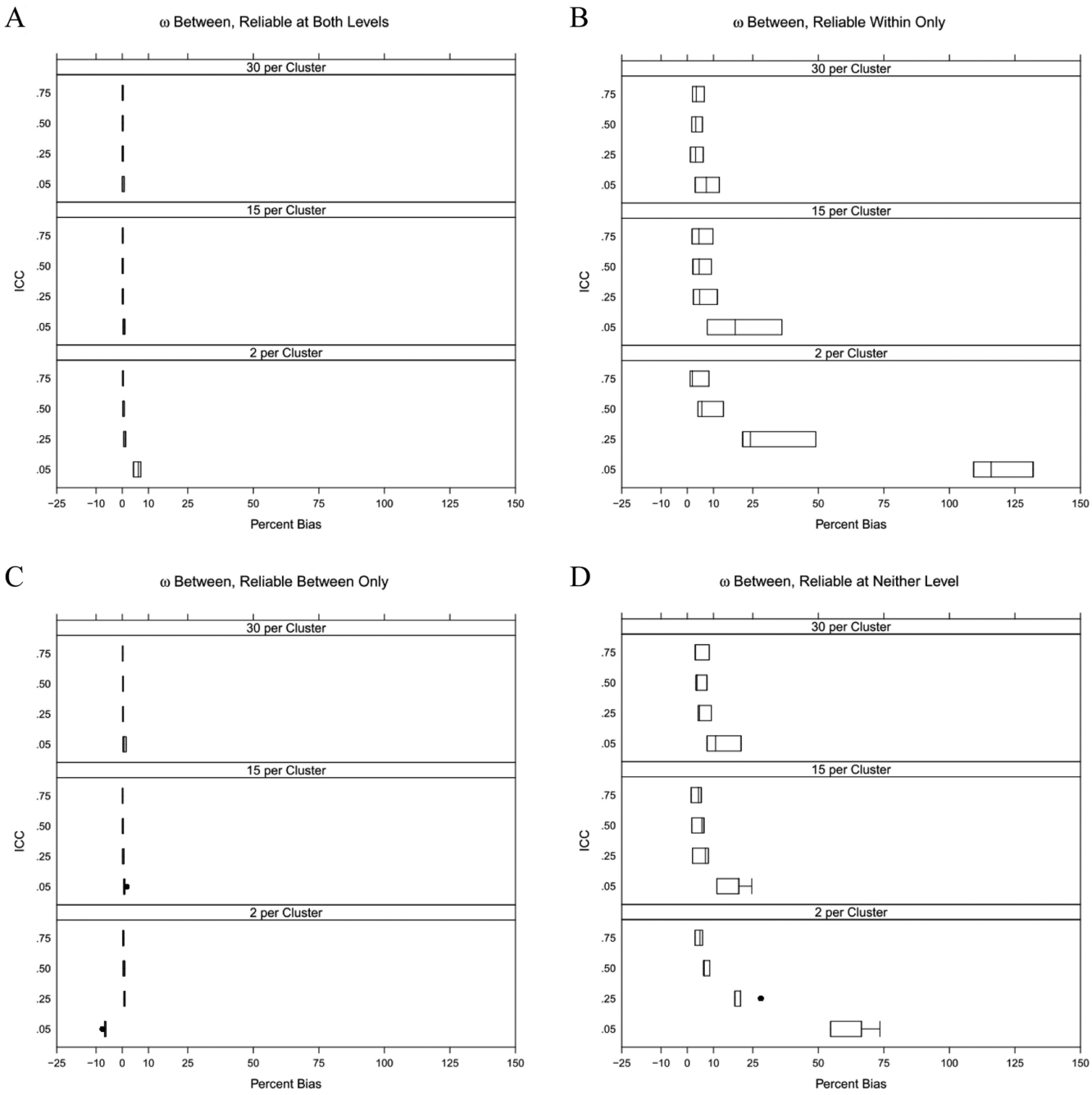

Figure 6. Bias of between-cluster ω. ICC = intraclass correlation.

dict the performance of H (Hypothesis 3B). Our results generally support Hypothesis 3A, with within-level α and ω both displaying low levels of bias. The exception was that within-level ω was biased when the data were grouped into very small clusters (i.e., dyads). The same three-way interaction predicted between-cluster α and ω. The uniformly positive bias for between-cluster ω when ICC, cluster size, and between-cluster reliability were small was not as immediately apparent for between-cluster α, however. Instead, between-cluster α also displayed a tendency to display negative bias when ICC, cluster size, and within-level reliability all were low.

Our Hypothesis 3B, that the performance of H would be influenced by different factors than the other two reliability estimates considered, received less consistent support. Within-level H appeared to be biased by the same factors as within-level ω, although between-cluster H was biased in almost every condition for which actual reliability at the between level was low.

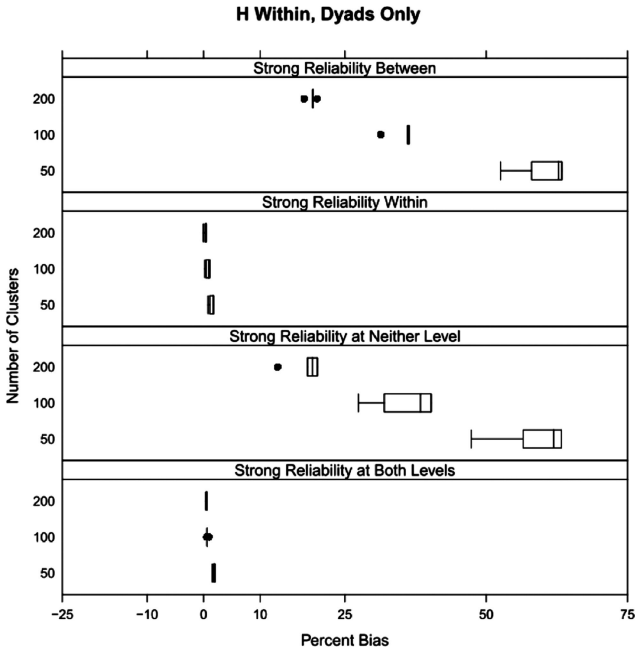

Figure 7. Bias of within-level H.

Convergence Rates

Models for computing single- and multilevel α did not present convergence problems, supporting Hypothesis 4A. At least 90% of all models converged in every condition for single-level α and for all conditions computing two-level α except when ICC = .05, there were two observations per cluster, and the scale had low reliability at both levels (83.4% convergence) or was reliable at only the between level (89.2% convergence). Our two-level CFA models displayed a wider range of convergence rates, however.[8]

Convergence rates for the two-level CFA models were predicted by all possible two-way interactions between the reliability condition, the number of clusters, and the number of observations per cluster, $F(26, 125164) = 60{,}560.00$, $p < .001$, $\eta^2 = .926$. As Figure 9 shows, convergence was especially low when there were few observations per cluster, fewer clusters, and reliability at the within level was low. These results partially support Hypotheses 4B and 4C, indicating that convergence was negatively impacted by sample size at both levels and by population reliabilities.

Monte Carlo Confidence Intervals

Monte Carlo and empirical 95% confidence intervals are presented for both α and ω in Tables 3 and 4. Due to the large amount of information presented in each table, these tables present results for only two sample sizes that represent large and small samples, respectively. Due to the especially poor performance of between-cluster H and the resultant fact that we cannot endorse its use in applied multilevel analyses, confidence intervals for H are not provided.

Results show close concordance between the empirical and Monte Carlo confidence limits when reliability was high. The exception from this trend was that our calculator underestimated the lower confidence limit and overestimated the upper confidence limit for reliability at the between level when ICCs were low and within-level reliability was also low. These results conditionally support our Hypothesis 5, suggesting that confidence limits from our Monte Carlo calculator closely matched the empirical confidence limits of our level-specific reliability estimates as long as level-specific reliability was not low.

When level-specific reliability was low, however, lower confidence limits tended to be negatively biased, especially when ICCs were low. These results therefore suggest that our calculator performs best when the lower confidence limit is reasonably high (e.g., greater than .50) but may underestimate lower confidence limits when the lower bound is small, especially when sample sizes are also small. As we discuss below, the issue of small sample sizes may be specific to studies characterized by a small number of cases (e.g., five or fewer observations) per cluster.

Credible Intervals

Tables 5 and 6 present 95% Bayesian credible intervals and matching empirical quantiles for the same conditions described when testing the Monte Carlo confidence limits above. As the tables show, the 95% credible intervals were biased in many of the same conditions as Monte Carlo confidence limits, but the amount of bias tended to be far less severe. The upper limit of the estimated credible interval provided an unbiased estimate of the 97.5th quantile of the observed sampling distribution for both α and ω in all conditions, with only one exception. Lower credible limits also tended to provide unbiased estimates of the 2.5th empirical percentiles, with all instances of bias occurring when the empirical 2.5th percentile was low (i.e., less than .30). Whereas the lower limits were not systematically biased in any direction, these results nevertheless suggest difficulty estimating small lower bounds. As such, we can strongly recommend Bayesian credible intervals with the caveat that lower limits may be misestimated when the lower end of the actual sampling distribution is very low (i.e., less than .30), thus supporting our Hypothesis 6.

Applied Example

We next present an empirical example involving multilevel reliability estimation using data from the 2007 Trends in International Mathematics and Science Study (TIMSS; Williams et al., 2009). Analyses included 7,475 children nested in 515

[8] Convergence rates specify the number of models that both (a) converged and (b) displayed minimally acceptable model fit (i.e., RMSEA < .08, CFI > .90, and TLI > .90). Of the models that converged, less than 10% displayed poor model fit except for the single-level CFA model. Approximately 24% of these models displayed poor fit, likely because the data generating model and the fitted model were markedly different. We removed cases with poor model fit under the assumption that applied users would reject these models. As such, our discussion of bias speaks directly to those instances where an analyst might reasonably attempt to estimate a construct's reliability.

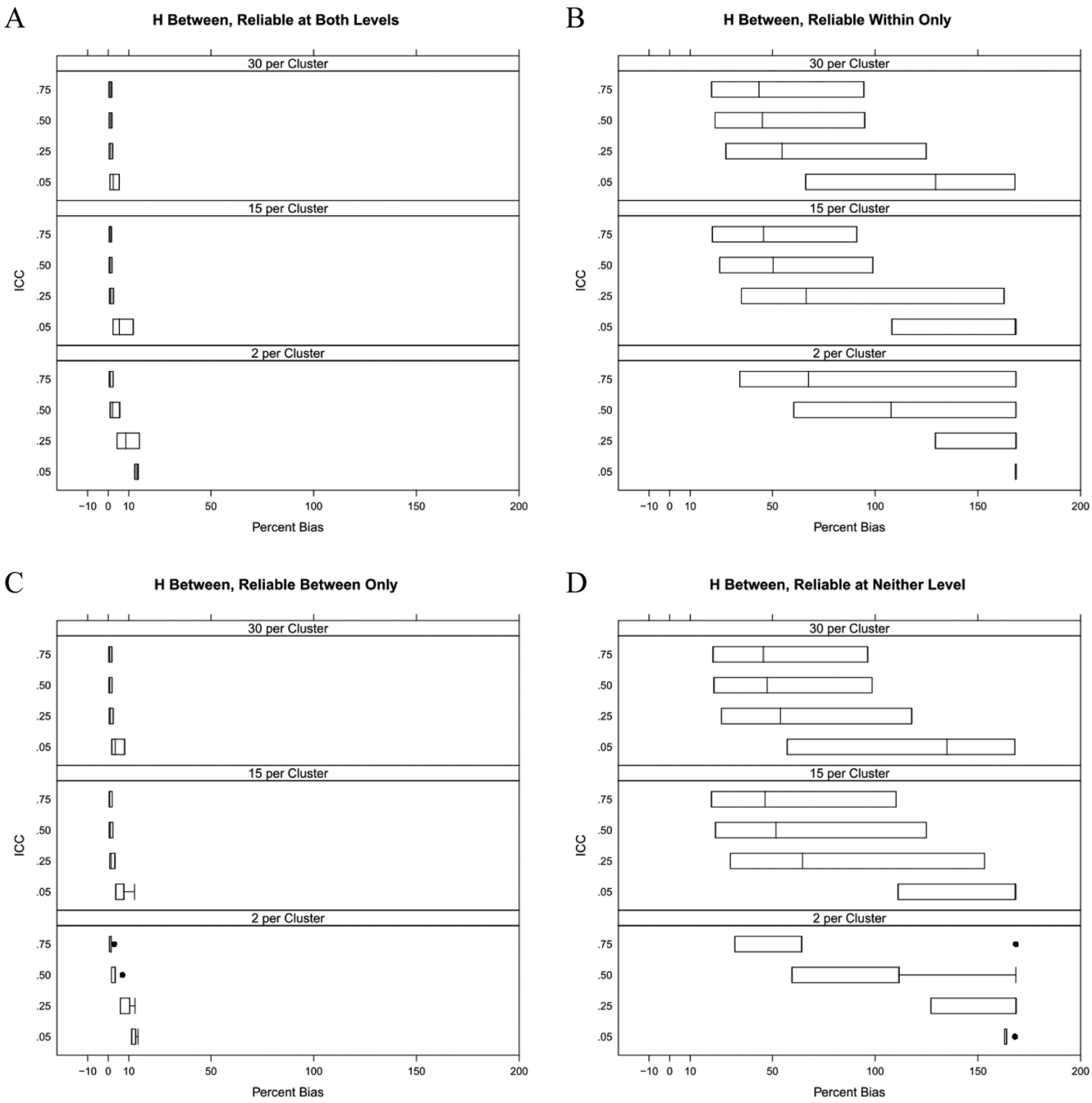

Figure 8. Bias of between-cluster H. ICC = intraclass correlation.

schools with complete data[9] for four items representing fourth graders' positive attitudes toward math (TIMSS variables: AS4MAMOR, AS4MAENJ, AS4MALIK, and AS4MABOR [reverse-coded]). These items were selected because they show strong interitem correlations (all *r*s > .50) and displayed sufficient variability at the child and classroom levels to justify multilevel modeling (i.e., all ICCs > .05). Mplus code for this example is presented in Appendix C in the online supplemental materials.

Single-level α was estimated by specifying a saturated single-level covariance structure with the variance of a unit-weighted composite and α both included as model parameters. In the Mplus syntax, the scale score's composite variance was specified as

[9] Children with missing data on any of these four variables were omitted from this example.

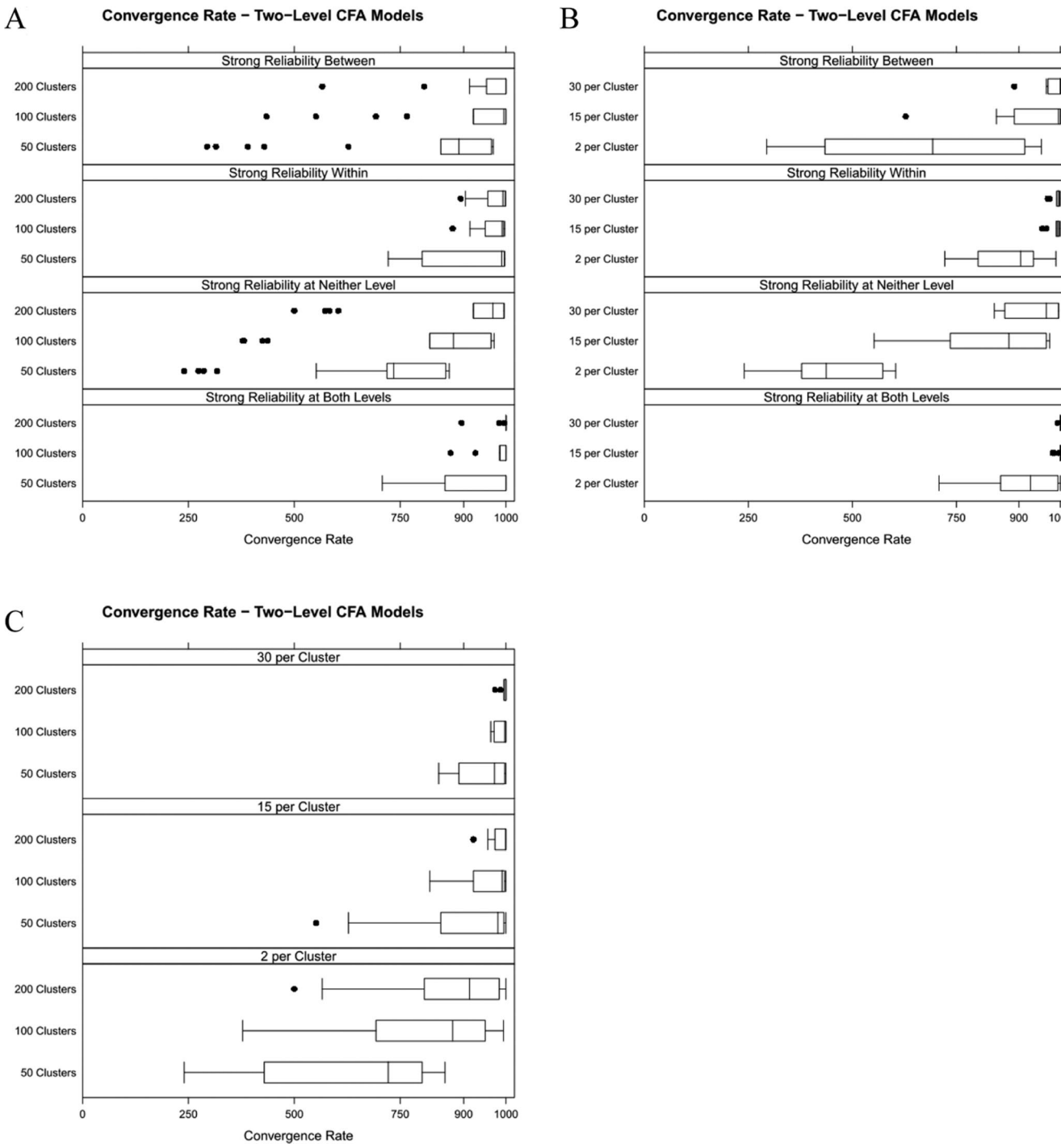

Figure 9. Convergence rates for the two-level CFA models. A: The interaction between the number of clusters and the reliability condition. B: The interaction between cluster size and the reliability condition. C: The interaction between the number of clusters and cluster size. CFA = confirmatory factor analysis.

COMP_V = V1 + V2 + V3 + V4 + 2*(C1 + C2 + C3 + C4 + C5 + C6),

where V1–V4 represent each indicator's variance and C1–C6 represent the six item covariances. The item covariances and composite score variance were then used to compute α. The Mplus syntax for α is

ALPHA = (((C1 + C2 + C3 + C4 + C5 + C6)/6)*16)/ COMP_V,

Table 3
Empirical Versus Monte Carlo Confidence Intervals—200 Clusters, 30 Observations per Cluster

	High ICC						Low ICC					
	Monte Carlo		Empirical		% bias[a]		Monte Carlo		Empirical		% bias[a]	
Estimate	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
						High reliability within only						
Within												
α	0.8455	0.8569	0.8451	0.8574	0.04	−0.06	0.8453	0.8573	0.8451	0.8574	0.02	−0.02
ω	0.8479	0.8592	0.8474	0.8595	0.06	−0.04	0.8478	0.8595	0.8475	0.8595	0.03	−0.01
Between												
α	0.2150	0.4952	0.2019	0.4979	6.47	−0.53	0.0165	0.5944	−0.1611	0.6104	***−110.24***	−2.63
ω	0.2292	0.5045	0.2271	0.5075	0.91	−0.59	0.0740	0.6130	0.0221	0.6151	***235.45***	−0.34
						High reliability between only						
Within												
α	0.3473	0.3955	0.3463	0.3971	0.26	−0.40	0.3486	0.3954	0.3463	0.3971	0.66	−0.42
ω	0.3477	0.3960	0.3470	0.3974	0.20	−0.36	0.3490	0.3957	0.3470	0.3974	0.58	−0.42
Between												
α	0.8131	0.8799	0.8147	0.8809	−0.20	−0.11	0.7942	0.8861	0.7957	0.8911	−0.19	−0.57
ω	0.8173	0.8824	0.8183	0.8837	−0.13	−0.14	0.8020	0.8899	0.8017	0.8946	0.04	−0.53

Note. ICC = intraclass correlation.
[a] Percent bias greater than 10% in bold italics.

where 16 in the numerator represents the number of items (four) squared. The model resulted in α = .868, suggesting good overall reliability.

A similar two-level model was then fit to the same data, with additional parameters separately estimating the variance of a unit-weighted composite and α for both the within and between levels. Results displayed acceptable reliability at each level and indicated that the scale was slightly more reliable between groups (α = .974, 95% CI [.960, .984]) than within groups (α = .856, 95% CI [.848, .863]).

We next fit a single-level unidimensional CFA, with both ω and H estimated as model parameters using model-estimated factor loadings and residual variances (see Equations 3, 5, and 6). Mplus syntax for both reliability estimates is

```
NUM = (L1 + L2 + L3 + L4)**2,

DENOM = ((L1 + L2 + L3 + L4)**2) + (R1 + R2 + R3 +
        R4),

OMEGA = NUM/DENOM,

H = 1/(1 + (1/((L1**2/R1) + (L2**2/R2) + (L3**2/R3) +
    (L4**2/R4)))),
```

Table 4
Empirical Versus Monte Carlo Confidence Intervals—100 Clusters, Two Observations per Cluster

	High ICC						Low ICC					
	Monte Carlo		Empirical		% bias[a]		Monte Carlo		Empirical		% bias[a]	
Estimate	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
						High reliability within only						
Within												
α	0.7988	0.8831	0.8076	0.8832	−1.09	−0.02	0.7964	0.8875	0.8101	0.8751	−1.69	1.41
ω	0.8037	0.8869	0.8161	0.8862	−1.52	0.07	0.8023	0.8904	0.8170	0.8781	−1.80	1.40
Between												
α	−0.0957	0.6468	−0.4689	0.6085	***−79.60***	6.29	−11.151	0.9452	0.1109	0.9052	***−101.52***	4.43
ω	0.0195	0.6777	0.0388	0.6137	***−49.68***	***10.42***	0.0205	0.9998	0.2906	0.9415	***−92.96***	6.20
						High reliability between only						
Within												
α	0.1375	0.5446	0.1805	0.5070	***−23.82***	7.42	0.1425	0.5423	0.2357	0.5062	***−39.53***	7.13
ω	0.1635	0.5579	0.2710	0.5201	***−39.66***	7.27	0.1671	0.5564	0.2972	0.5032	***−43.76***	***10.57***
Between												
α	0.7805	0.8948	0.7851	0.8974	−0.58	−0.29	−1.6090	1.1091	−0.2774	0.8419	***479.97***	***31.74***
ω	0.7902	0.9004	0.8028	0.8976	−1.57	0.31	0.0168	1.5200	0.4635	0.9203	***−96.37***	***65.16***

Note. ICC = intraclass correlation.
[a] Percent bias greater than 10% in bold italics.

Table 5
Empirical Versus Bayesian Credible Intervals—200 Clusters, 30 Observations per Cluster

	High ICC						Low ICC					
	Credible interval		Empirical		% bias[a]		Credible interval		Empirical		% bias[a]	
Estimate	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
						High reliability within only						
Within												
α	0.8440	0.8570	0.8451	0.8574	−0.13	−0.05	0.8460	0.8560	0.8451	0.8574	0.11	−0.16
ω	0.8480	0.8610	0.8474	0.8595	0.07	0.18	0.8480	0.8600	0.8475	0.8595	0.06	0.05
Between												
α	0.2330	0.5170	0.2019	0.4979	***15.38***	3.84	−0.2630	0.6180	−0.1611	0.6104	***63.24***	1.24
ω	0.1970	0.4950	0.2271	0.5075	***−13.25***	−2.46	0.1250	0.5550	0.0221	0.6151	***466.79***	−9.78
						High reliability between only						
Within												
α	0.3360	0.3950	0.3463	0.3971	−2.99	−0.53	0.3430	0.3980	0.3463	0.3971	−0.97	0.23
ω	0.3440	0.3940	0.3470	0.3974	−0.86	−0.85	0.3460	0.3970	0.3470	0.3974	−0.29	−0.10
Between												
α	0.8120	0.8770	0.8147	0.8809	−0.34	−0.45	0.7750	0.8880	0.7957	0.8911	−2.60	−0.35
ω	0.8150	0.8860	0.8183	0.8837	−0.41	0.26	0.7940	0.8920	0.8017	0.8946	−0.96	−0.29

Note. ICC = intraclass correlation.
[a] Percent bias greater than 10% in bold italics.

where L1–L4 represent the four unstandardized factor loadings and R1–R4 represent the estimated residual variances. Results roughly matched those for α, with both ω and H suggesting slightly higher reliability than α (ω = .868, H = .892).

The same model was then fit as a two-level CFA with a congeneric factor structure specified at each level. Initial estimation produced a negative residual variance for item AS4MAENJ at the between level, and all residual variances were constrained to remain greater than zero when obtaining final two-level reliability estimates. Results for the two-level model match those for two-level α, with both ω and H showing greater reliability between groups (ω = .977, 95% CI [.964, .987]; H = .999, 95% CI not provided) than within groups (ω = .857, 95% CI [.849, .863]; H = .882, 95% CI not provided). These very high estimates suggest that the indicators do not substantially differ at the between level, such that the interitem between-cluster correlations are near unity. All items strongly reflect the between-cluster factor, although this result says little about the between-cluster factor's validity. In this instance, any single indicator modeled at the between level would be as informative as a between-cluster latent construct. In such a circumstance, a researcher could justifiably simplify the between-cluster model by considering only a single indicator.

The between level estimates further highlight the importance of the single strongest factor loading when calculating H, which in this example had a standardized loading very close to 1.00. In other words, this example underscores the point that esti-

Table 6
Empirical Versus Bayesian Credible Intervals—100 Clusters, Two Observations per Cluster

	High ICC						Low ICC					
	Credible interval		Empirical		% bias[a]		Credible interval		Empirical		% bias[a]	
Estimate	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
						High reliability within only						
Within												
α	0.7890	0.8860	0.8076	0.8832	−2.31	0.31	0.8000	0.8810	0.8101	0.8751	−1.24	0.67
ω	0.8070	0.8970	0.8161	0.8862	−1.12	1.22	0.8160	0.8870	0.8170	0.8781	−0.12	1.02
Between												
α	−0.1240	0.6620	−0.4689	0.6085	***−73.56***	8.79	0.5550	0.8810	0.1109	0.9052	***400.31***	−2.67
ω	0.0000	0.5800	0.0388	0.6137	***−100.00***	−5.49	0.0050	0.9050	0.2906	0.9415	***−98.28***	−3.87
						High reliability between only						
Within												
α	0.1830	0.5270	0.1805	0.5070	1.38	3.95	0.2280	0.5370	0.2357	0.5062	−3.26	6.08
ω	0.1490	0.5410	0.2710	0.5201	***−45.03***	4.02	0.2310	0.5420	0.2972	0.5032	***−22.27***	7.70
Between												
α	0.7740	0.9090	0.7851	0.8974	−1.41	1.29	−0.4590	0.7770	−0.2774	0.8419	***65.45***	−7.71
ω	0.7830	0.9080	0.8028	0.8976	−2.47	1.16	0.0010	0.7990	0.4635	0.9203	***−99.78***	***−13.18***

Note. ICC = intraclass correlation.
[a] Percent bias greater than 10% in bold italics.

mates of H will be high as long as there is at least one strong factor loading, regardless of whether or not H is high or low in the population.

Discussion

The consequences of ignoring a data set's hierarchical structure have been thoroughly explored in the hypothesis-testing literature, but the need to account for clustering when estimating reliability has been largely ignored. In this article, we have extended three common reliability estimates to a multilevel framework using MCFA, advocating the estimation of level-specific reliability when dealing with multilevel data. Our simulations show that within-level reliability estimates are generally unbiased although positive bias can arise for ω when within-level reliability is low and there are relatively few, small clusters. Between-cluster reliability estimates were also unbiased under most data conditions, but ω displayed positive bias when item ICCs were low, especially when the between-cluster reliability was also low and when there were few observations per cluster. Between-cluster α behaved similarly but displayed negative bias when within-level reliability was low for dyadic data. These results generally support our argument for computing level-specific reliability when researchers are faced with multilevel data. We can draw two additional conclusions from our results. First, between-cluster maximal reliability was biased in nearly every condition characterized by low reliability at the between level. Maximal reliability is bounded by the reliability of the strongest indicator and may be unduly biased when even a single factor loading is overestimated. Given the propensity for H to overestimate its population value at the between level, we cannot recommend its use in empirical multilevel research.

A second theme in our results was the tendency to see bias at the between level, as well as convergence difficulties, for conditions with few observations per cluster, low ICCs, and low reliability at the between level. Although this result can be partially supported by the simple fact that there was limited information at the between level under these conditions, it is also possible that these results represent the larger problem of dealing with negative ICCs in the multilevel modeling framework. Multilevel models, including MCFA and MSEM, necessarily constrain ICCs to positive values, as evidenced by the fact that ICC is often taken as the proportion of between-cluster variance in an item, relative to its total variance in a sample. ICC can alternatively (and more accurately) be discussed as the expected correlation between pairs of Level 1 units sampled from a given cluster, however, making it possible to have negative ICCs (termed *negative nonindependence*; Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Random samples drawn from populations in which the between-cluster covariance matrix is nearly nonpositive definite, as would occur when few observations per cluster represent between-cluster data with low ICCs and low reliabilities, would be especially prone to containing negative expected covariances between same-cluster pairs. The result would be a negative ICC, leading to either model misfit or nonconvergence. Further investigation is needed, however, to confirm whether this possibility indeed induced bias or model nonconvergence in our simulation.

Suggestions for Researchers

The above discussion indicates that level-specific reliability estimates (i.e., α and ω) are generally preferable to single-level estimates whenever ICCs are nontrivial (i.e., ≥.05). Within-level estimates may be untrustworthy when clusters are small, however, especially in instances of dyadic data. Under dyadic data conditions, our results suggest that within-level α is preferable to within-level ω, despite the fact that α is an inconsistent estimator of reliability in the population. Due to the inconsistency of α, our results suggest that within-level ω is preferred in all other data conditions.

The issue of small clusters is especially relevant to researchers who examine repeated-measures data using a multilevel modeling framework and when there are very few observations per individual (i.e., few Level 1 units).[10] We can anticipate from our simulation that within-level reliability estimates should be relatively unbiased under such conditions, but Level 2 reliability estimates cannot be trusted when ICCs are low. Despite this obvious limitation, we recommend a multilevel approach to estimating reliability over alternative methods when dealing with a small number of Level 1 units. Two common alternatives would be to either (a) report reliability for each wave separately while ignoring the nested structure of the data set or (b) take a fixed-effects approach that only corrects standard errors (e.g., using the Huber-White correction). These alternatives necessarily conflate reliability across levels and will be biased to the extent that scale reliability differs across levels and the item ICCs are greater than zero.

Similarly, between-cluster ω is preferable to between-cluster α under most data conditions, although we suggest between-cluster α for data with small ICCs and smaller clusters (i.e., 15 or fewer observations per cluster). Due to model convergence rates, between-cluster α may also serve as a generally acceptable fallback for instances in which a two-level CFA fails to converge.

Last, our Monte Carlo confidence limit calculator produced unbiased limits when reliability was high but biased limits when reliability was low. Until more work is done to determine why our confidence intervals tended to be biased under certain conditions, we can recommend our calculator for computing confidence intervals only when estimated reliability is greater than .60, as this value fell roughly at (or above) the upper limit of the empirical 95% confidence intervals for conditions in which our calculator performed poorly. Bayesian credible limits provided less-biased estimates, however, and we can give a much stronger recommendation for their use in a broad range of data conditions. As mentioned in our results, the one caveat is that lower (but not upper) credible interval limits can misrepresent the actual lower quantiles of empirical sampling distributions when the empirical lower quantiles are smaller than .30. Furthermore, we estimated credible intervals using only noninformative priors, so our results do not necessarily generalize to cases where informative priors would be more appropriate.

[10] But see Huang and Weng (2012) for a discussion of level-specific reliability in the context of many repeated measures as occurs in ecological momentary assessment data.

Limitations and future directions. Our simulation results provide initial guidance to researchers who wish to estimate multilevel reliability, but additional work on the topic could be useful. We limited our discussion to single-factor MCFA models with fixed (i.e., nonrandom) factor loadings at Level 1. We therefore do not account for noncongeneric scales, models for which the factor structure varies across levels, nonnormal (e.g., binary) data, the analysis of tetrachoric or polychoric correlations, or models where Level 1 reliability is allowed to vary across Level 2 units. A great deal of research has extended methods for single-level reliability estimation, and future work should similarly examine such extensions to multilevel reliability estimation. Furthermore, our simulation data were both generated and analyzed using a single software package (Mplus), and it is difficult to determine the impact of using a single program on our results. Future research should confirm the previous results using separate software packages. Despite these limitations, our article clearly underscores the benefits of multilevel reliability estimation and highlights conditions for which multilevel reliability estimates may be biased.

References

Bentler, P. M. (2007). Covariance structure models for maximal reliability of unit-weighted composites. In S. Lee (Ed.), Handbook of computing and statistics with applications*: Vol. 1. Handbook of latent variable and related models* (pp. 1–19). New York, NY: Elsevier.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels: A typology of composition models. *Journal of Applied Psychology, 83,* 234–246. doi:10.1037/0021-9010.83.2.234

Conger, A. J. (1980). Maximally reliable composites for unidimensional measures. *Educational and Psychological Measurement, 40,* 367–375. doi:10.1177/001316448004000213

Connor, C. M., Ponitz, C. C., Phillips, B. M., Travis, Q. M., Glasney, S., & Morrison, F. J. (2010). First graders' literacy and self-regulation gains: The effect of individualizing student instruction. *Journal of School Psychology, 48,* 433–455. doi:10.1016/j.jsp.2010.06.003

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin, 32,* 917–929. doi:10.1177/0146167206287721

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. doi:10.1007/BF02310555

de Gruijter, D. N. M., & van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioral sciences*. New York, NY: Chapman & Hall.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11,* 189–228. doi:10.1214/ss/1032280214

Efron, B. (2011). The bootstrap and Markov chain Monte Carlo. *Journal of Biopharmaceutical Statistics, 21,* 1052–1062. doi:10.1080/10543406.2011.607736

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis* (2nd ed.). New York: Wiley.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16,* 625–641. doi:10.1080/10705510903203573

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Robers (Eds.), *Handbook of advanced multilevel analysis* (pp. 163–172). New York, NY: Routledge.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467. doi:10.1007/BF02294400

Gottfredson, N. A., Panter, A. T., Daye, C. E., Allen, W. F., & Wightman, L. F. (2009). The effects of educational diversity on a national sample of law students: Fitting multilevel latent variable models in data with categorical indicators. *Multivariate Behavioral Research, 44,* 305–331. doi:10.1080/00273170902949719

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255–282. doi:10.1007/BF02288892

Halaby, C. N. (2004). Panel models in sociological research: Theory intro practice. *Annual Review of Sociology, 30,* 507–544. doi:10.1146/annurev.soc.30.012703.110629

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future—A festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL: Scientific Software International.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6,* 87–93.

Huang, P.-H., & Weng, L.-J. (2012). Estimating the reliability of aggregated and within-person centered scores in ecological momentary assessment. *Multivariate Behavioral Research, 47,* 421–441. doi:10.1080/00273171.2012.673924

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83,* 126–137.

Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika, 69,* 459–474. doi:10.1007/BF02295646

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika, 2,* 151–160. doi:10.1007/BF02288391

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika, 62,* 245–249. doi:10.1007/BF02295278

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

MacCallum, R. C., Widaman, K. F., Preacher, K., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36,* 611–637. doi:10.1207/S15327906MBR3604_06

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84–99. doi:10.1037/1082-989X.4.1.84

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39,* 99–128. doi:10.1207/s15327906mbr3901_4

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology, 23,* 1–21. doi:10.1111/j.2044-8317.1970.tb00432.x

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

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Survey Research Methods, 3,* 45–58.

Muthén, B. O. (1990). *Mean and covariance structure analysis of hierarchical data* (UCLA Statistics Series No. 62). Retrieved from http://gseis.ucla.edu/faculty/muthen/articles/Article_032.pdf

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22,* 376–398. doi:10.1177/0049124194022003006

Muthén, B. O., & Asparouhov, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 15–40). New York, NY: Taylor & Francis.

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462. doi:10.1007/BF02294365

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32,* 1–13. doi:10.1007/BF02289400

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6,* 77–98. doi:10.1080/19312458.2012.679848

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21,* 173–184. doi:10.1177/01466216970212006

Raykov, T. (1998). A method for obtaining standard errors and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement, 22,* 369–374. doi:10.1177/014662169802200406

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research, 37,* 89–103. doi:10.1207/S15327906MBR3701_04

Raykov, T. (2004). Estimation of maximal reliability: A note on a covariance structure modeling approach. *British Journal of Mathematical and Statistical Psychology, 57,* 21–27. doi:10.1348/000711004849295

Raykov, T., & du Toit, S. H. C. (2005). Estimation of reliability for multiple-component measuring instruments in hierarchical designs. *Structural Equation Modeling, 12,* 536–550. doi:10.1207/s15328007sem1204_2

Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling, 13,* 130–141. doi:10.1207/s15328007sem1301_7

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Taylor & Francis.

Raykov, T., & Penev, S. (2010). Evaluation of reliability coefficients for two-level models via latent variable analysis. *Structural Equation Modeling, 17,* 629–641. doi:10.1080/10705511.2010.510052

Raykov, T., & Shrout, P. E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling, 9,* 195–212. doi:10.1207/S15328007SEM0902_3

Shavelson, R. J., & Webb, N. M. (2004). Generalizability theory. In K. Kemp-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 99–105). Oxford, England: Elsevier.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. doi:10.1007/s11336-008-9101-0

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Thomson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology, 30,* 357–366.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21,* 81–90.

Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34,* 25–33. doi:10.1177/001316447403400104

Wilhelm, P., & Schoebi, D. (2007). Assessing mood in daily life: Structural validity, sensitivity to change, and reliability of a short-scale to measure three basic dimensions of mood. *European Journal of Psychological Assessment, 23,* 258–267. doi:10.1027/1015-5759.23.4.258

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., . . . Stearns, P. (2009). TIMSS 2007 U.S. Technical report and user guide (NCES 2009-012). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Woolridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70,* 123–133. doi:10.1007/s11336-003-0974-7

Received April 6, 2012
Revision received December 18, 2012
Accepted February 6, 2013 ■

Correction to Franić, Dolan, Borsboom, Hudziak, van Beijsterveldt, and Boomsma (2013)

In the article "Can Genetics Help Psychometrics? Improving Dimensionality Assessment Through Genetic Factor Modeling" by Sanja Franić, Conor V. Dolan, Denny Borsboom, James J. Hudziak, Catherina E. M. van Beijsterveldt, and Dorret I. Boomsma (*Psychological Methods*, Vol. 18, No. 3, pp. 406–433. doi: 10.1037/a0032755), funding information was omitted from the author note. The author note should have stated that this research was funded by the European Research Council Grant 230374 to D.I. Boomsma.

DOI: 10.1037/a0036139