Composite Reliability of Multilevel Data: It's About Observed Scores and Construct Meanings

Mark H. C. Lai

University of Southern California

Author Note

Mark H. C. Lai, Department of Psychology, University of Southern California

I would like to thank Oi-man Kwok and Hio Wa Mak for the constructive feedback on an early version of this manuscript.

Correspondence concerning this article should be addressed to Mark Lai (Email: hokchiol@usc.edu), Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061.

Abstract

This paper shows how the concept of reliability of composite scores, as defined in classical test theory, can be extended to the context of multilevel modeling. In particular, it discusses the contributions and limitations of the various level-specific reliability indices proposed by Geldhof, Preacher, and Zyphur (2014), denoted as $\tilde{\omega}^b$ and $\tilde{\omega}^w$ (and also $\tilde{\alpha}^b$ and $\tilde{\alpha}^w$). One major limitation of those indices is that they are quantities for latent, unobserved level-specific composite scores, and are not suitable for observed composites at different levels. As illustrated using simulated data in this paper, $\tilde{\omega}^b$ can drastically overestimate the true reliability of between-level composite scores (i.e., observed cluster means). Another limitation is that the development of those indices did not consider the recent conceptual development on construct meanings in multilevel modeling (Stapleton & Johnson, 2019; Stapleton, Yang, & Hancock, 2016). To address the second limitation, this paper defines reliability indices (ω^{2l}, ω^b, ω^w, α^{2l}, α^b, α^w) for three types of multilevel observed composite scores measuring various multilevel constructs: individual, configural, shared, and within-cluster. The paper also shows how researchers can obtain sample point and interval estimates using the derived formulas and the provided R and Mplus code. In addition, a large-scale national data set was used to illustrate the proposed methods for estimating reliability for the three types of multilevel composite scores, and practical recommendations on when different indices should be reported are provided.

Keywords: multilevel, reliability, composite, alpha, omega

Composite Reliability of Multilevel Data: It's About Observed Scores and Construct Meanings

Psychological and social science researchers commonly deal with data with a multilevel structure, such as students nested within schools and survey participants nested within neighborhoods. Before analyzing multilevel data with multi-item psychological instruments, researchers should ensure that their instruments have good score reliability by obtaining and reporting reliability information of their instruments based on their sample data (Appelbaum et al., 2018). However, for data of multilevel nature, it is often less straightforward to compute such information. Geldhof, Preacher, and Zyphur (2014) pointed out that previous studies on reliability of multilevel data generally conflated reliability within groups and between groups, and have proposed several statistics to estimate within-level and between-level composite reliability estimates using multilevel confirmatory factor analysis (MCFA), which I denoted as $\tilde{\omega}^b$ and $\tilde{\omega}^w$ for between-level and within-level composite reliability, and $\tilde{\alpha}^b$ and $\tilde{\alpha}^w$ for between-level and within-level generalizations of Cronbach's α (Cronbach, 1951). Since then, many researchers have adopted their methods.[1]

However, as to be discussed in this paper, there are two major issues with $\tilde{\omega}^b$ and $\tilde{\omega}^w$ (and $\tilde{\alpha}^b$ and $\tilde{\alpha}^w$). First, in classical test theory, reliability is a property of *observed* test scores (e.g., Lord, 1955; Thompson, 2003), but the definition of between-level reliability (i.e., $\tilde{\omega}^b$ in Geldhof et al. (2014) is a property of unobserved, *latent* cluster means, which ignores the sampling error in the *observed* cluster means (see Lüdtke, Marsh, Robitzsch, & Trautwein, 2011, for a detailed discussion). As a result, and as shown in Jak, Oort, and Dolan (2014), when strong factorial invariance (Millsap, 2011) across clusters holds, meaning that the measurement intercepts and factor loadings are the same across all clusters, the $\tilde{\omega}^b$ estimate is always 1.0 regardless of how unstable the between-level observed composite scores are, and can mislead researchers to think that the scores of the instrument has very good or perfect reliability. In the present paper, I demonstrate that $\tilde{\omega}^b$ proposed in Geldhof et al. can be substantially larger than the true reliability of observed between-level composite scores (i.e., the cluster means of individual composite scores in a cluster, as explained later in the paper). I also show analytically under what conditions of

number of clusters, cluster sizes, and intraclass correlations would $\tilde{\omega}^b$ be inflated most when estimating the reliability of the between-level observed composite scores.

Second, the use of $\tilde{\omega}^b$ and $\tilde{\omega}^w$ implicitly assumes that there are two completely separate constructs at the within and between levels, measured by two sets of composite scores. However, in conventional multilevel modeling, researchers sometimes use only the overall composite (i.e., composite of raw item scores), and in other times decompose the overall composite into the within-level and between-level components. In addition, Stapleton et al. (2016) has recently identified different conceptualizations of constructs in multilevel data with different construct meanings, but to my knowledge there has not been discussions on how multilevel composite reliability fits into their framework. In this paper, I propose methods to compute sample reliability coefficients for the three types of multilevel composite scores, respectively under three measurement model specifications with different construct meanings (i.e., individual and configural constructs, shared construct, and within-cluster construct). I then demonstrate the calculations of three proposed indices using data from the 2007 Trends in International Mathematics and Science Study (TIMSS; Williams et al., 2009). Finally, I discuss similar issues with $\tilde{\alpha}^b$ and $\tilde{\alpha}^w$, and propose similar extensions of single-level α to the three types of multilevel composite scores.

Multilevel Factor Model

To simplify the discussion without loss of generality, I limit the scope to an instrument of p items measuring one latent construct at both the within and between levels. With slight differences from the notations used in Geldhof et al. (2014), a multilevel factor model is defined as

$$\mathbf{Y}_{ij} = \boldsymbol{\kappa} + \boldsymbol{\lambda}_j^w \eta_{ij}^w + \boldsymbol{\lambda}^b \eta_j^b + \zeta_{ij}^w + \zeta_j^b, \tag{1}$$

where $\mathbf{Y}_{ij}$ is a vector of p measured variables of the ith individual in the jth cluster, η_{ij}^w and η_j^b are his or her latent variable scores at the within and the between level, respectively, $\boldsymbol{\kappa}$ is a vector of p measurement intercept, $\boldsymbol{\lambda}_j^w$ is a vector of p within-level factor loadings for cluster j, $\boldsymbol{\lambda}^b$ is a vector

of p between-level factor loadings, and $\boldsymbol{\zeta}_j^b$ and $\boldsymbol{\zeta}_{ij}^w$ are vectors of measurement errors for the p items at the within and the between levels, respectively. The random variables η^w, η^b, $\boldsymbol{\zeta}^w$, and $\boldsymbol{\zeta}^b$ are assumed mutually independent with $\eta_{ij}^w \sim \mathcal{N}(0, \phi^w)$, $\eta_j^b \sim \mathcal{N}(0, \phi^b)$, $\zeta_{ij}^w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_j^w)$, $\zeta_j^b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}^b)$. Researchers commonly assume local independence such that $\boldsymbol{\Theta}_j^w = \text{diag}[\theta_{11}^w \dots \theta_{pp}^w]$ and $\boldsymbol{\Theta}^b = \text{diag}[\theta_{11}^b \dots \theta_{pp}^b]$, but composite reliability can still be computed when such an assumption is violated, as later discussed in this paper. Following Geldhof et al. I also assume $\boldsymbol{\lambda}_j^w = \boldsymbol{\lambda}^w$ (i.e., equal factor loadings across clusters) and $\boldsymbol{\Theta}_j^w = \boldsymbol{\Theta}^w$ (homogeneity of error covariances across clusters) for all j. Geldhof et al. also made the cross-level loading equality assumption that $\boldsymbol{\lambda}^w = \boldsymbol{\lambda}^b$ (see also Jak et al., 2014). A path diagram of the model is shown in Figure 1.

In classical test theory, reliability is a property of observed scores, as also pointed out in Geldhof et al. (2014, see also Lord, 1955; Lord & Novick, 1968), so we need to first define the observed composite scores with multilevel data. Specifically, the overall composite of p items for person i in cluster j is simply the sum of the p item scores:

$$Z_{ij} = \sum_{k=1}^{p} Y_{ijk}, \tag{2}$$

where k indexes items. It is defined in the same way as its counterpart in single-level data. In addition, with multilevel data, researchers commonly perform cluster-mean centering (Enders & Tofighi, 2007) to disentangle the composite score into the within-level and between-level components. The between-level observed composite score is

$$Z_j^b = \sum_{i=1}^{n_j} Z_{ij}/n_j, \tag{3}$$

where n_j is the cluster size for j. The within-level composite score is the deviation of the individual composite score from the group mean, that is,

$$Z_{ij}^w = Z_{ij} - Z_j^b. \tag{4}$$

It should be emphasized that all three of Z_{ij}, Z_j^b, and Z_{ij}^w are observed scores, and the decomposition of Z_{ij} into the within-level and the between-level components is the same as the usual practice of cluster mean centering in multilevel analysis (Enders & Tofighi, 2007). As an example, a researcher may be interested in using job satisfaction, operationalized as a composite of multiple self-report items, to predict job performance for employees nested within companies, and examine whether company-level job satisfaction (Z_j^b) and individual-level job satisfaction (Z_{ij}^w) predict job performance differently. The difference between the two prediction coefficients is commonly denoted as the contextual effect.

Geldhof et al. (2014) defined several between-level and within-level reliability coefficients, including variants of the traditional coefficient α, composite reliability ω, and maximal reliability H (Conger, 1980; Hancock & Mueller, 2001). Because they recommended the use of between-level $\tilde{\omega}^b$ and within-level $\tilde{\omega}^w$ as they had the best performance in their simulation and are more in line with the multilevel factor model, in the present paper I mainly focus on multilevel ω coefficients, but also note towards the end how the discussion applies to α. I do not discuss H given that it refers to reliability of weighted composites that are different from the ones presented in (2), (3), and (4), and is beyond the scope of the present paper.[2] I denote the within-level and between-level ω coefficients proposed in Geldhof et al. as $\tilde{\omega}^b$ and $\tilde{\omega}^w$ here to distinguish them from the indices proposed in this paper, where

$$\tilde{\omega}^b = \frac{(\sum_{k=1}^{p} \lambda_k^b)^2}{(\sum_{k=1}^{p} \lambda_k^b)^2 + \sum_{k=1}^{p} \theta_{kk}^b} \tag{5}$$

$$\tilde{\omega}^w = \frac{(\sum_{k=1}^{p} \lambda_k^w)^2}{(\sum_{k=1}^{p} \lambda_k^w)^2 + \sum_{k=1}^{p} \theta_{kk}^w}. \tag{6}$$

The above equations assume that η_j^b and η_{ij}^w have been standardized such that $\phi^b = \phi^w = 1$ and that local independence holds at both levels. When local independence is violated, the unique factor covariances need to be added to the denominator for both equations (5) and (6). Without the assumption of local independence and the condition that ϕ^b and/or ϕ^w are standardized,

equations (5) and (6) become:

$$\tilde{\omega}^b = \frac{(\sum_{k=1}^{p} \lambda_k^b)^2 \phi^b}{(\sum_{k=1}^{p} \lambda_k^b)^2 \phi^b + \mathbf{1'\Theta}^b\mathbf{1}}, \tag{7}$$

$$\tilde{\omega}^w = \frac{(\sum_{k=1}^{p} \lambda_k^w)^2 \phi^w}{(\sum_{k=1}^{p} \lambda_k^w)^2 \phi^w + \mathbf{1'\Theta}^w\mathbf{1}}. \tag{8}$$

where $\mathbf{1'\Theta}^b\mathbf{1} = \sum_{k=1}^{p}\sum_{k'=1}^{p} \theta_{kk'}^b$ and $\mathbf{1'\Theta}^w\mathbf{1} = \sum_{k=1}^{p}\sum_{k'=1}^{p} \theta_{kk'}^w$ are the sums of all elements in $\mathbf{\Theta}^b$ and $\mathbf{\Theta}^w$, respectively. As can be seen, $\tilde{\omega}^b$ represents the proportion of variance explained by η^b at the between level, and $\tilde{\omega}^w$ represents the proportion of variance explained by η^w at the within level.

Although the definitions of $\tilde{\omega}^w$ and $\tilde{\omega}^b$ may seem straightforward and intuitive, I identify two issues that make these indices potentially problematic in practice, namely (a) that they are properties of the composite of latent, unobserved variables at the between-level and within-level, but in classical test theory, reliability is a property of observed composite scores, and (b) that they do not take into account the different types of multilevel constructs with different meanings as discussed in Stapleton et al. (2016). Below I detail the two issues, and propose alternative indices that both are reliability indices of observed composite scores and fit into Stapleton et al.'s framework.

Issue I: $\tilde{\omega}^b$ and $\tilde{\omega}^w$ Are Reliability Indices of Unobserved, Latent Scores

As pointed out earlier, in classical test theory, reliability is a property of observed test score, not of any latent or abstract construct. This point was also discussed in Geldhof et al. (2014). Indeed, in single-level analysis, ω is called composite reliability because it was derived as the reliability of the observed composite score—the unweighted sum of item scores. It is defined as the proportion of variance of the composite score attributable to the true score.

In classical test theory, an observed test score Y_i for person i is decomposed as

$$Y_i = T_i + e_i \tag{9}$$

where T_i is the expected value of Y_i, also called the true score, and e_i is the random error component. Reliability is then defined as the squared correlation between Y and T, which is also equal to the variance of T divided by the variance of Y (Lord & Novick, 1968, Chapter 3).

Therefore, to understand $\tilde{\omega}^b$ and $\tilde{\omega}^w$, it is important to know what their corresponding scores are. The derivation in Geldhof et al. (2014) implicitly uses latent mean centering (Asparouhov & Muthén, 2019; Lüdtke et al., 2008) in the multilevel structural equation modeling framework by defining

$$\mathbf{Y}_{ij} = \mathbf{Y}_{j}^{b(l)} + \mathbf{Y}_{ij}^{w(l)}, \tag{10}$$

$$\mathbf{Y}_{j}^{b(l)} = \boldsymbol{\kappa} + \boldsymbol{\lambda}^{b}\eta_{j}^{b} + \zeta_{j}^{b}, \tag{11}$$

$$\mathbf{Y}_{ij}^{w(l)} = \boldsymbol{\lambda}_{j}^{w}\eta_{ij}^{w} + \zeta_{ij}^{w}, \tag{12}$$

where $\mathbf{Y}_{j}^{b(l)} = [Y_{j1}^{b(l)} \cdots Y_{jp}^{b(l)}]$ contains the latent, error-free item means of cluster j, and $\mathbf{Y}_{ij}^{w(l)} = [Y_{ij1}^{w(l)} \cdots Y_{ijp}^{b(l)}]$ contains the deviation scores of individual i in cluster j from the latent cluster means. The (l) superscript is used to make clear that $\mathbf{Y}^{b(l)}$ and $\mathbf{Y}^{w(l)}$ are latent, unobserved variables. Thus, they are represented by circles in Figure 1. Under this representation, and from equations (7) and (8), it is clear that $\tilde{\omega}^b$ is the composite reliability of the sum score of the latent cluster means of p items, $Z_{j}^{b(l)} = \sum_{k=1}^{p} Y_{jk}^{b(l)}$, and $\tilde{\omega}^w$ is the composite reliability of the sum score of the deviation scores of p items from the latent cluster means, $Z_{ij}^{w(l)} = \sum_{k=1}^{p} Y_{ijk}^{w(l)}$.

The problem is, of course, that $Z^{b(l)}$ and $Z^{w(l)}$ are not observed variables, whereas in classical test theory, the focus is on decomposing the *observed scores* into components of true scores and random errors (Lord & Novick, 1968). Indeed, in single-level analysis, composite reliability ω (as well as α) is meaningful because it represents the proportion of true score variance in the observed composite scores (Raykov, 1997; Thompson, 2003), and the observed composite scores are the imperfect scores that test users can obtain for assessment, classification, and other research purposes. Otherwise, if one computes the reliability of an unobserved score, such as the true score T in equation (9), one very likely obtains an unrealistically high reliability.

In the case of the true score T, because there is no random error in it, the reliability of T is always perfect but meaningless, because we cannot directly obtain T. The same is true for multilevel data: reliability should be defined for observed but not latent scores.

Consider the observed composite, Z_j^b, and the latent composite, $Z_j^{b(l)}$, at the between level. In practice, unless one has an infinitely large cluster size, the observed sample mean of item k in a cluster of size n_j, $\bar{Y}_{jk} = \sum_{i=1}^{n_j} Y_{ijk}/n_j$, will be different from the latent cluster mean $Y_{jk}^{b(l)}$. For example, if Z_j^b represents the mean job satisfaction score of a sample of employees in a company, it is only an estimate of and is not the same as $Z_j^{b(l)}$, the true mean of the company, unless the sample contains everyone in the company. The sampling error of Z_j^b has a variance of $\text{Var}(Y_{jk}^{w(l)})/n_j$. Therefore, Z_j^b, which is the unweighted sum of the sample cluster means across all items, has two sources of error, namely measurement error of the items at the between level (i.e., θ^b), as well as the sampling error corresponding to the difference between Z_j^b and $Z_j^{b(l)}$. These two sources of error were thoroughly discussed in Lüdtke et al. (2011, see also Raykov & Marcoulides, 2006).[3] Because $\tilde{\omega}^b$ only captures one source of error, generally it overestimates the true composite reliability of Z_j^b. The following simulation provides further evidence for that.

Illustration of the Bias of $\tilde{\omega}^b$

To illustrate the bias of $\tilde{\omega}^b$, I simulated a large sample based on the model shown in Figure 1 and defined in (10) to (12), with five items, 10,000 clusters of size 10, $\phi^w = 1$, $\phi^b = 0.25$, mean of $\eta = 0$, and for all items $\kappa = 0$, $\lambda^b = \lambda^w = 0.5$, $\theta^w = 1$, $\theta^b = 0.1$. I set the number of clusters to be large so that the parameter and reliability estimates are close to the population values. The simulation code is included in the supplemental material. In the data generation process, I first simulated the unobserved η^b values for each cluster and the η^w values for each individual, which are used to compute $\mathbf{Y}^b$ and $\mathbf{Y}^w$. The observed variable scores are then obtained as $\mathbf{Y} = \mathbf{Y}^b + \mathbf{Y}^w$. I also computed the observed cluster means, $\bar{Y}_{j1}, \ldots, \bar{Y}_{jp}$, for each item and for every cluster j. Using the definition of reliability from classical test theory, I computed the squared correlation of the between-level observed composite, Z^b, with the true latent factor score, that is, $[\text{Corr}(Z^b, \eta^b)]^2$,

as the true reliability of Z^b. I also computed $\tilde{\omega}^b$ as defined in Geldhof et al. (2014). If $\tilde{\omega}^b$ measures the reliability of the observed composite Z^b, it should be close to $[\text{Corr}(Z^b, \eta^b)]^2$; however, given the previous discussion, I expected $\tilde{\omega}^b$ to be larger than $[\text{Corr}(Z^b, \eta^b)]^2$, because of the substantial sampling error in the observed cluster means due to a small cluster size of 10.

Using the simulated data set, $[\text{Corr}(Z^b, \eta^b)]^2 = .490$, which is the true reliability of the observed composite of cluster means. However, the $\tilde{\omega}^b$ estimate was .756, which overestimated the true reliability of Z^b by 54%. On the other hand, the squared correlation between the latent composite and the true latent factor score at the between level is $[\text{Corr}(Z^{b(l)}, \eta^b)]^2 = .756$, confirming that $\tilde{\omega}^b$ estimates the reliability of the composite score of *latent* cluster means. Therefore, $\tilde{\omega}^b$ can be highly misleading when used to justify the measurement quality of between-level composite scores in multilevel analysis, especially when the cluster size is small.

Consequences of overestimated reliability. Because $\tilde{\omega}^b$ can drastically overestimate the reliability at the between level, it potentially leads to inflated reliability information of the measurement in published research that involved multilevel data. For example, using $\tilde{\omega}^w$ and $\tilde{\omega}^b$, Rush and Hofer (2014) reported that the score reliability coefficients of positive and negative affects were .80 to .84 at the within-person level, but were much higher at .94 to .97 at the between-person level. The inflated $\tilde{\omega}^b$ values may give researchers a false sense of confidence in their measurement when doing multilevel analyses.

The inflated between-level reliability indicated by $\tilde{\omega}^b$ or $\tilde{\alpha}^b$ was perhaps most problematic in scale development and validation. For example, in the development of the Instructional Skills Questionnaire, Knol, Dolan, Mellenbergh, and van der Maas (2016) presented only the teacher-level $\tilde{\alpha}^b$, which were between .90 and .99, in the main text as evidence for reliability, and presented the student-level $\tilde{\alpha}^w$, which were between .49 and .79, in supplemental material. Although the authors were correct to use a teacher-level reliability coefficient as the instrument is intended to give teacher-level scores as mean scores across student raters, the inflated $\tilde{\alpha}^b$ values made the reliability of the teacher scores looked much better than they were. For example, using equation (19) to be discussed in a later section, the teacher-level composite scores from the

Instruction subscale in Knol et al. (2016), which had $\tilde{\alpha}^b$ reported as .90, only had a true α^b of .76 (assuming an ICC of .06 as reported in the paper), after taking into account the sampling error for a class size of 82. If another researcher uses the same subscale to measure instruction quality based on a class size of 25, the expected reliability of the teacher scores will only be .56, which says that the scores are much less stable than what was presented as $\tilde{\alpha}^b$ = .90.

Issue II: $\tilde{\omega}^b$ and $\tilde{\omega}^w$ Do Not Take Into Account Different Kinds of Multilevel Constructs

Based on the seminal work by Kozlowski and Klein (2000) and Marsh et al. (2012), Stapleton et al. (2016) and Stapleton and Johnson (2019) identified four possible types of constructs in a multilevel CFA setting, including (a) an individual construct, where the construct is defined at the within level but possibly has nonzero intraclass correlations across clusters (e.g., social skill of an individual), (b) a configural construct, also called a contextual construct by Marsh et al., which consists of cluster averages of individual constructs (e.g., mean student achievement of a school), (c) a shared construct, also called a climate construct by Marsh et al., where a construct is defined purely at the between level and is inherently a characteristic of a cluster (e.g., school climate), and (d) a within-cluster construct, where the comparison of a construct is only meaningful within a cluster but not across clusters (e.g., sociometric ratings within a classroom). These four types of constructs, and the corresponding MCFA specification, are discussed below, followed by a discussion of the proposed methods to compute reliability for the three composite scores (i.e., Z_{ij}, Z_j^b, Z_{ij}^w). The corresponding R (R Core Team, 2019) and Mplus (L. K. Muthén & Muthén, 2017) code for computing the proposed reliability indices can be found in the supplemental material.

Individual Construct

Individual constructs are generally the most common in psychological and behavioral research. An example MCFA model of an individual construct with four items, decomposed into $\mathbf{Y}^{b(l)}$ and $\mathbf{Y}^{w(l)}$, is shown in Figure 1. Note that under the MCFA framework, even though there is only one individual construct, the distribution of the latent variable, η, will be decomposed into

the between-level component, η^b, and the within-level component, η^w. If η is an observed variable, this decomposition is no different from the usual practice of decomposing a variable into the between-level and within-level components in multilevel modeling (Asparouhov & Muthén, 2019; Enders & Tofighi, 2007). In such a model, η_j^b is the population cluster mean for cluster j, and η_{ij}^w is the individual-level deviation of person i from the cluster mean. By the definition of an individual construct, the factor loadings need to be constrained to be equal (i.e., $\boldsymbol{\lambda}^b = \boldsymbol{\lambda}^w$) so that η^b and η^w are on the same metric (Jak et al., 2014; Stapleton et al., 2016), and the intraclass correlation of η is $\eta^b/(\eta^b + \eta^w)$ (Mehta & Neale, 2005). The residual variances and covariances among $Y^{b(l)}$ reflect either additional shared construct at the cluster level (Stapleton et al., 2016) or violations of measurement invariance across clusters (Jak et al., 2014).[4]

Configural Construct

Stapleton and Johnson (2019) referred to η_j^b in Figure 1, which consists of the latent cluster means of the individual construct η, as a configural construct. Although such a distinction of the various components of a latent variable was only recently made, the practice of cluster-mean centering has been the standard in traditional multilevel modeling. When researchers perform cluster-mean centering, it is commonly expected that the between-level and the within-level components may have differential associations with other variables. A classic example is the "big-fish-little-pond" effect (Marsh, 1987), where it was found the association between student-level ability and academic self-concept was generally positive, but the association between school-average ability and school-average self-concept was generally negative and smaller. As a result, with cluster-mean centering, the two components of an individual construct are treated as two separate constructs, and a configural construct refers to the between-level component.

Reliability of composites measuring individual and configural constructs. Because individual and configural constructs are part of the same model, I discuss the reliability of the corresponding composites together based on the model in Figure 1. When the overall composite (Z_{ij}) is used to measure an individual construct, such as social skill and IQ, because the scores can

be compared both within a cluster and across clusters, the true score variance should include both $\text{Var}(\eta^b) = \phi^b$ and $\text{Var}(\eta^w) = \phi^w$; in other words, the true score variance is $\text{Var}(\eta) = \phi^b + \phi^w$. Therefore, the composite reliability of Z_{ij} measuring an individual construct, ω^{2l}, is

$$\omega^{2l} = \frac{(\sum_{k=1}^{p} \lambda_k)^2(\phi^w + \phi^b)}{(\sum_{k=1}^{p} \lambda_k)^2(\phi^w + \phi^b) + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}}. \tag{13}$$

Note that equation (13) is essentially identical to the two-level composite reliability derived in Raykov and du Toit (2005, p. 540, equation 10). Compared to $\tilde{\omega}^w$ in equation (8), ω^{2l} captures the population variance of both the within-level and the between-level components of the true score and of the errors.

In addition, for an individual construct, when cluster-mean centering is used to decompose the overall composite into the between-level composite (Z_j^b) and the within-level composite (Z_{ij}^w), reliability should be computed for these two composites. For the composite deviation score Z_{ij}^w, the variance is

$$\text{Var}(Z_{ij}^w) = \frac{n_j - 1}{n_j}\left[\left(\sum_{k=1}^{p} \lambda_k\right)^2 \phi^w + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}\right]. \tag{14}$$

Therefore, the true score variance component is $[(n_j - 1)/n_j](\sum_{k=1}^{p} \lambda_k^w)^2\phi^w$. When defining the reliability of Z_{ij}^w as the true score variance divided by $\text{Var}(Z_{ij}^w)$, the constant $(n_j - 1)/n_j$ will be canceled out. Therefore, the reliability of Z_{ij}^w is

$$\omega^w = \frac{(\sum_{k=1}^{p} \lambda_k)^2\phi^w}{(\sum_{k=1}^{p} \lambda_k)^2\phi^w + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}}, \tag{15}$$

which is the same as $\tilde{\omega}^w$ as defined in equation (8).

For the reliability of Z_j^b, the between-level composite, first note that observations within a cluster share the same latent cluster means $\mathbf{Y}^{b(l)}$ but are conditionally independent. Thus, within a cluster j and for $i \neq i'$, $\text{Cov}(\mathbf{Y}_{ij}, \mathbf{Y}_{i'j}) = \text{Var}(\mathbf{Y}^{b(l)}) = \phi^b\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Theta}^b$, whereas the variance of each observation is $\text{Var}(\mathbf{Y}_{ij}) = \text{Var}(\mathbf{Y}^{b(l)}) = (\phi^b + \phi^w)\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Theta}^b + \boldsymbol{\Theta}^w$. This implies that the covariance matrix of the observed cluster item means is $\text{Var}(\bar{\mathbf{Y}}_j) = (\phi^b + \phi^w/n_j)\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Theta}^b + \boldsymbol{\Theta}^w/n_j$, which

depends on the cluster size n_j (see also Raudenbush & Bryk, 2002, Chapter 3). Therefore, the variance of Z_j^b is

$$\text{Var}(Z_j^b) = \left(\sum_{k=1}^{p} \lambda_k\right)^2 (\phi^b + \phi^w/n_j) + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}/n_j, \tag{16}$$

where $(\sum_{k=1}^{p} \lambda_k)^2\phi^w/n_j + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}/n_j$ is the sampling error variance of the observed cluster means. Assuming equal cluster size such that $n_j = n$ for all j, the reliability of the between-level composite is

$$\omega^b = \frac{(\sum_{k=1}^{p} \lambda_k)^2\phi^b}{(\sum_{k=1}^{p} \lambda_k)^2(\phi^b + \phi^w/n) + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}/n}. \tag{17}$$

When cluster sizes are not equal, the reliability will be different for clusters with different cluster sizes. In that case, one way to define an overall measure of reliability is to first obtain the mean variance of Z^b as $\text{E}[\text{Var}(Z^b)] = (\sum_{k=1}^{p} \lambda_k)^2(\phi^b + \phi^w)/\tilde{n} + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}/\tilde{n}$, where $\tilde{n} = 1/(\sum_{j=1}^{J} 1/n_j)$ is the harmonic mean of cluster sizes. Then, the between-level composite reliability can be obtained by simply replacing n_j in equation (17) by $\tilde{n}$. Note that this use of harmonic mean is consistent with previous literature in multilevel modeling when considering the reliability of cluster means of observed variables (e.g., Kwok et al., 2008).

It should be pointed out that although the between-level and the within-level components of an individual construct are parts of the same construct, the two parts nevertheless may have differential relationships with other variables. Therefore, the two components of an individual construct are usually treated as two separate variables, in which case both ω^b and ω^w should be reported. Furthermore, even when a researcher only uses the overall composite without decomposition for his or her research, future researchers may use the same instrument with decomposition into the between-level and within-level components. Therefore, to facilitate comparisons across research, I recommend always reporting ω^{2l}, ω^b, and ω^w for an individual construct with multilevel data, especially for research involving scale development.

Relationships between ω^{2l} and $\tilde{\omega}$ ($\tilde{\omega}^b$ and $\tilde{\omega}^w$). To more precisely understand the relationship between $\tilde{\omega}^w$, $\tilde{\omega}^b$, and ω^{2l} for overall composites measuring individual constructs, one

should note that from equations (7) and (8), equation (13) can be expressed as:

$$\omega^{2l} = (1 - \rho_Z)\tilde{\omega}^w + \rho_Z\tilde{\omega}^b, \tag{18}$$

where $\rho_Z = \text{Var}(Z^b)/\text{Var}(Z)$ is the intraclass correlation of the observed composite Z. Therefore, the true reliability of the overall composite scores of an individual construct is a weighted average of $\tilde{\omega}^w$ and $\tilde{\omega}^b$, and $\tilde{\omega}^w = \omega^{2l}$ only when $\rho_Z = 0$ or when $\tilde{\omega}^b = \tilde{\omega}^w$. Otherwise, the bias of $\tilde{\omega}^w$ as an estimator of ω^{2l} is $[\rho_Z/(1 - \rho_Z)](\omega^{2l} - \omega^b)$, which is larger for larger ρ_Z and larger difference between ω^{2l} and $\tilde{\omega}^b$. For example, when $\omega^{2l} = .8$, with a small ICC of .05 for Z and a low $\tilde{\omega}^b$ of .60, the bias is only .01; however, with a larger ICC of .40 for Z and a perfect $\tilde{\omega}^b$ of 1.0 (which happens when strong factorial invariance across clusters holds, as previously discussed), ω^w will be .67, which severely underestimates the reliability of an observed composite of an individual construct.

Similarly, to understand the bias when using $\tilde{\omega}^b$ to approximate ω^b, note that equation (17) can be expressed as:

$$\omega^b = \tilde{\omega}^b \frac{\rho_Z}{\rho_Z + (1 - \rho_Z)/\tilde{n}}. \tag{19}$$

Because the largest possible value of $\tilde{\omega}^b$ is 1.0, the largest possible value of ω^b is $\rho_Z/[\rho_Z + (1 - \rho_Z)/\tilde{n}]$, which can be much smaller than 1.0 when cluster size is small. For example, when $\rho_Z = .3$, $\tilde{n} = .5$, the maximum of ω^b is $.3/(.3 + .7/5) = .68$. The bias of $\tilde{\omega}^b$ as an estimator of ω^b is $(1 - \rho_Z)\tilde{\omega}^b/\tilde{n}/[\rho_Z + (1 + \rho_Z)/\tilde{n}]$, which is zero only when $\tilde{\omega}^b = 0$ (i.e., when there are no true score variance), when $\rho_Z = 1$ (i.e., when there is no level-1 variability), or when $\tilde{n}$ is a huge number such that $\tilde{n} \to \infty$. Therefore, whereas $\tilde{\omega}^b$ can be an approximation of the reliability of the between-level composite when the cluster size is large, or a theoretical upper bound of ω^b, it generally gives an inflated reliability coefficient, especially with smaller ρ_Z and smaller $\tilde{n}$.

A special case is when strong invariance across clusters holds, which, as discussed in Jak et al. (2014), will always result in $\tilde{\omega}^b = 1$, regardless of how small the cluster size and the ICC of Z is. If one is interested in using the between-level composite to measure a configural construct such

as school-level achievement, but the scores across students in a school are highly spread out, then one should expect to get very different group means when different samples of students are surveyed. In that case, the actual reliability of the group-level composite should be low, but one can still get $\tilde{\omega}^b = 1$.

Shared Construct

Based on the discussion in Stapleton et al. (2016), an example MCFA model of a shared construct with four items is shown in Figure 2. What makes a shared construct different from a configural construct is that it is inherently an attribute of a cluster but is measured by multiple informants at the within level. Examples include safety of a neighborhood, effectiveness of a teacher, and organizational climate (e.g., Liao & Chuang, 2007). The teacher-level constructs measured by the Instructional Skills Questionnaire (Knol et al., 2016) presented earlier are also shared constructs. In these situations, within-level variations are not true score variance at the between level, so a saturated within-level model is specified with $\text{Cov}(\mathbf{Y}^{w(l)}) = \Sigma^w$, and a unidimensional factor model is specified at the between level to reflect the shared construct.

Reliability of composites measuring a shared construct. For a shared construct measured by the between-level composite, Z_j^b as defined in equation (3), the reliability ω^b is similarly computed as in (17) for the model with an individual and a configural construct, except that now the within level is a saturated model. The corresponding composite reliability for Z_j^b measuring a shared construct is thus

$$\omega^b = \frac{(\sum_{k=1}^{p} \lambda_k^b)^2 \phi^b}{(\sum_{k=1}^{p} \lambda_k^b)^2 \phi^b + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\Sigma^w\mathbf{1}/\tilde{n}}, \tag{20}$$

As discussed in Stapleton and Johnson (2019), it is also possible that items intended to measure a shared construct also measure an individual construct at the within level, resulting in an additional configural component at the between level. In this situation, the reliability coefficient in (20) for the between-level composite reflects its consistency for simultaneously measuring a shared and a configural constructs. However, researchers may be more interested in the proportion

of true score variance due to only the shared construct, in which case the simultaneous shared-and-configural model proposed in Stapleton et al. (2016) and Stapleton and Johnson should be used. In the Appendix I presented an example and extensions of ω for a shared construct under this simultaneous shared-and-configural model.

Within-Cluster Construct

An example MCFA model of a within-cluster construct with four items is shown in Figure 3. Here, the latent variable is only meaningful at the within level and is denoted as η^w. An example of a within-cluster construct is students' popularity among their peers within the same classroom based on some sociometric ratings, such that each student is only compared to other students in the same classroom, but not to others in different classrooms. In this case, any variations across classrooms, if any, should be irrelevant to the construct. Thus, a saturated between-level model is used.

Reliability of composites measuring a within-cluster construct. For a within-level composite, Z_{ij}^w, measuring a within-cluster construct, the composite reliability is

$$\omega^w = \frac{(\sum_{k=1}^{p} \lambda_k^w)^2 \phi^w}{(\sum_{k=1}^{p} \lambda_k^w)^2 \phi^w + \mathbf{1'\Theta}^w\mathbf{1}}, \tag{21}$$

which is essentially the same as (15), except that the factor loadings are specific to the within level as the between-level model is a saturated one.

Empirical Illustration

To illustrate the differences between the two $\tilde{\omega}$ coefficients and the three ω coefficients presented in this paper, I revisit the applied example analysis from the 2007 Trends in International Mathematics and Science Study (TIMSS Williams et al., 2009) as described in Geldhof et al. (2014). In the example, 7,896 students from 515 schools in the United States responded to four items on attitudes toward math (`AS4MAMOR`: Would like to do more math, `AS4MAENJ`: I enjoy learning mathematics, `AS4MALIK`: I like math, and reverse coded `AS4MABOR`:

Math is boring). The items were scored on a 4-point scale from 1-*agree a lot* to 4-*disagree a lot*.[5] There were 2.1% to 3.7% of missing data on the four items; following Geldhof et al. (2014) I only analyzed data with scores on all four items, resulting in an analytic sample size of 7,475. Table 1 shows the within-level and between-level interitem correlations and covariances based on a saturated multilevel structural equation modeling model using maximum likelihood estimation in `lavaan` (Rosseel, 2012).

Using the level-specific indices by Ryu and West (2009), a one-factor model at the within level with a saturated between-level model produced acceptable fit, $\chi^2(df = 2, N = 7,475) = 6.11$, $p = .047$, RMSEA = .017, 95% CI [.002, .032], CFI = 1.00. At the between-level an initial one-factor model resulted in a small but negative ($\theta = -0.001$) unique factor variance estimate for `AS4MAEN`; given the small value, I fixed it to zero. The fit of the one-factor between-level with a saturated within-level model produced reasonable fit, $\chi^2(df = 3, N = 7,475) = 6.18$, $p = .103$, RMSEA = .064, 95% CI [0.00, 0.14], CFI = .988. An MCFA with equal factor loadings across levels showed good fit, with overall $\chi^2(df = 8, N = 7,475) = 40.11$, $p < .001$, RMSEA = .023, 95% CI [.016, .031], CFI = .998; SRMR-within = .004 and SRMR-between = .054. However, likelihood ratio test showed that the loading constraints resulted in worse model fit, $\Delta\chi^2(df = 4) = 28.06$, $p < .001$, suggesting that there might be additional shared constructs at the between-level. Nevertheless, for illustrative purpose, I first computed reliability coefficients assuming unidimensionality at both levels for an individual construct.

The MCFA with equal loading constraints represents the specification of an individual construct with a configural component. The parameter estimates, using the `lavaan` syntax shown in the supplemental material, are given in Table 2. The estimated overall $\omega^{2l} = .867$ for the overall composite, 95% Wald CI [.862, .873], so the reliability was good for the overall composite score for measuring an individual-level constructs. For the within-level composite, $\tilde{\omega}^w = .861$, 95% CI [.855, .866], so in this example $\tilde{\omega}^w$ was slightly smaller than ω^{2l}. Finally, $\tilde{\omega}^b = .622$ for the between-level composite, 95% CI [.560, .684]. Therefore, the between-level composite was relatively low on reliability. In contrast, $\tilde{\omega}^b$, which ignores the sampling error due to within-level

variability, was estimated to be .975, 95% Wald CI [.964, .987]. Therefore, if one were to use $\tilde{\omega}^b$, one would mistakenly think that the observed composite of the cluster item means had close-to-perfect reliability, when the actual composite reliability was suboptimal in measuring a configural construct.

Because the four items mainly focus on students' attitudes towards math, with the students themselves being the subject of measurement, it is most natural that researchers would use the composite scores of the four items to measure an individual construct (with a configural component). However, for illustrative purpose, if the four items were to be used to measure a within-cluster construct that can only be compared within a cluster, meaning that one intends to compare students' attitudes toward math within each school, but not across schools, a saturated between-level model should be used. In this case, $\tilde{\omega}^w$ was similarly estimated to be .860, 95% Wald CI [.855, .865].

Because these four items mainly asked students' individual characteristics with no reference to intrinsic attributes of schools (e.g., teaching quality or school climate), they are not suitable for measuring shared constructs at the school level. An example involving the between-level composite reliability defined in (20) with items designed to measure a shared construct can be found in the Appendix, which also includes discussion on how reliability can be defined when the between-level composite measures both a shared and a configural construct.

Extension of Cronbach's Alpha (α) for Multilevel Composites of Different Construct Meanings

Here I briefly describe the extension of α to multilevel observed composites, which is largely similar to the discussion of ω. In single-level analyses, α (Cronbach, 1951) for a composite score Z is estimated by the ratio between the average covariance of the item scores and the total variance of the mean score. Assuming that the covariance matrix of p items is $\boldsymbol{\Sigma}$ with elements σ_{ij}, then α can be computed as

$$\alpha = \frac{p \sum_{k=2}^{p} \sum_{k'=1}^{p-1} \sigma_{kk'}}{(p-1)\mathbf{1}'\boldsymbol{\Sigma}\mathbf{1}}. \tag{22}$$

With multilevel data, similar to my previous discussion on ω^{2l}, α^{2l} for the overall observed composite can be obtained as

$$\alpha^{2l} = \frac{p \sum_{k=2}^{p} \sum_{k'=1}^{p-1} (\sigma_{kk'}^{w} + \sigma_{kk'}^{b})}{(p-1)[\mathbf{1}'\boldsymbol{\Sigma}^{b}\mathbf{1} + \mathbf{1}'\boldsymbol{\Sigma}^{w}\mathbf{1}]}. \tag{23}$$

It should be pointed out that α does not require estimations of parameters of a factor model, so its computation is the same whether a composite is used to measure a configural construct or a shared construct. For a between-level composite,

$$\alpha^{b} = \frac{p \sum_{k=2}^{p} \sum_{k'=1}^{p-1} \sigma_{kk'}^{b}}{(p-1)[\mathbf{1}'\boldsymbol{\Sigma}^{b}\mathbf{1} + \mathbf{1}'\boldsymbol{\Sigma}^{w}\mathbf{1}/\tilde{n}]}, \tag{24}$$

where $\tilde{n}$ is the harmonic mean of the cluster sizes. For a within-level composite,

$$\alpha^{w} = \frac{p \sum_{k=2}^{p} \sum_{k'=1}^{p-1} \sigma_{kk'}^{w}}{(p-1)\mathbf{1}'\boldsymbol{\Sigma}^{w}\mathbf{1}}. \tag{25}$$

When the essential tau-equivalence condition holds, which implies that the factor loadings are the same for all items, ω and α represents the same population quantities (assuming unidimensionality). It is thus not surprising that the relationships between the above α coefficients with those defined by Geldhof et al. (2014) (which I called $\tilde{\alpha}^{b}$ and $\tilde{\alpha}^{w}$) are similar to those between ω and $\tilde{\omega}$ coefficients as previously discussed. Note that point estimates of α^{2l}, α^{w}, and α^{b} only require estimates of the within-level and between-level item covariances, such as those presented in Table 1, but the corresponding standard error and interval estimates require additional information such as the asymptotic covariance matrix from a fully saturated MCFA model. For the TIMSS example, using the `lavaan` syntax shown in supplemental material, all α and $\tilde{\alpha}$ estimates were similar to the ω and $\tilde{\omega}$ estimates (see Table 3). Specifically, α^{2l} for the four attitudes toward math items was estimated to be .868, 95% Wald CI [.862, .873], which was slightly larger than the estimated $\tilde{\alpha}^{w}$ = .859, 95% Wald CI [.854, .865]. The estimate of α^{b} was .653, 95% Wald CI [.594, .711], which was much smaller than $\tilde{\alpha}^{b}$ = .968.

Notes on Confidence Intervals

The confidence intervals presented in the previous section were based on the asymptotic standard errors of the reliability coefficients using the delta method in standard SEM software. Geldhof et al. (2014) recommended two methods that should give more valid CIs with better coverage properties, namely the Monte Carlo CIs and the Bayesian credible intervals. I refer readers to Preacher and Selig (2012) for a discussion on the Monte Carlo CI and to B. Muthén and Asparouhov (2012) on Bayesian credible interval. For the TIMSS example, given the relatively large sample size, the different methods of computing interval estimates of multilevel reliability basically yielded the same results, as shown in Table 3. The supplemental material includes R codes and Mplus syntax for obtaining these two alternative types of interval estimates for the previous illustrative example.

Discussion

In this paper, I provide an updated perspective on how reliability should be defined for different kinds of constructs with multilevel data. In particular, I point out two limitations with the definitions of level-specific reliability provided in Geldhof et al. (2014), namely that they are measures of reliability of hypothetical, latent variables, but not observed variables, and that they do not consider different construct meanings that were recently proposed in Stapleton et al. (2016) and Stapleton and Johnson (2019). Recognizing such limitations in the existing literature on multilevel reliability, I provide reliability definitions for raw, between-level, and within-level composites when measuring various multilevel constructs: individual, configural, shared, and within cluster. These include extensions of the single-level composite reliability ω, a model-based index that does not require the essential tau equivalence assumption, to multilevel ω^{2l} for overall composite, ω^{b} for between-level composite, and ω^{w} for within-level composite, and extensions of the single-level Cronbach's α, which assumes essential tau equivalence but does not require fitting a factor model, to multilevel α^{2l}, α^{b}, and α^{w}. I have also showed how sample point and interval estimates of the six different reliability indices can be computed, provided reproducible R and

Mplus syntax, and illustrated the biases of the reliability statistics in Geldhof et al. (2014) analytically and through a simulated data set.

Reporting of reliability information is a basic requirement for empirical studies (Appelbaum et al., 2018; Flake, Pek, & Hehman, 2017), yet when it comes to multilevel constructs, Kim, Dedrick, Cao, and Ferron (2016) found that almost half (46%) of the studies they reviewed that used multilevel factor analysis did not report reliability information. I believe that the discussion of the six reliability coefficients for commonly obtained composite scores with multilevel data in the current paper will facilitate the reporting practices of such information in empirical multilevel studies. I urge researchers to consider the nature of their constructs, and report the corresponding reliability information as in single-level studies. In addition, just like any other point estimates, simply presenting ω or α estimates do not provide information about the uncertainty associated with those point estimates, so I also urge applied researchers report the corresponding confidence intervals using the procedures illustrated in the supplemental material.

I would like to emphasize again that when one should report ω^{2l}, ω^{b}, and/or ω^{w} (and similarly for the αs) depends on what types of composite scores are computed. For example, for an individual construct such as life satisfaction, if one uses the overall composite and does not center the scores or uses grand mean centering in multilevel analyses, one should compute ω^{2l}. However, if one uses group mean centering (with latent or observed group means; see Asparouhov & Muthén, 2019) to decompose the scores, which is generally the recommended procedure in multilevel modeling, one should compute both ω^{b} and ω^{w} as the between-level and within-level components are treated as two different variables and can have differential associations with other variables. To facilitate comparisons across studies and provide comprehensive evaluations of the measurements, a good strategy would be to always report all three of ω^{2l}, ω^{b}, and ω^{w} (or all three of α^{2l}, α^{b}, and α^{w}) for individual constructs.

Although the proposed reliability coefficients take into account different construct meanings and are properties of observed scores, several assumptions need to hold in order for them to be meaningful. As previously discussed, one assumption is that the latent factor variance is constant

across clusters. Although this homogeneity of variance assumption is regularly made in multilevel modeling, with MCFA the factor variance can be modeled as random as discussed in Stapleton et al. (2016), and with such a model the ω and α coefficients I discussed need to be adapted. Future research efforts can generalize the proposed reliability coefficients to account for heterogeneous factor variances. Similarly, the discussion in this paper assumes measurement invariance across clusters, which is also a testable assumption (Jak et al., 2014). When measurement parameters, such as factor loadings and unique factor variances, are allowed to vary across clusters, it implies different reliability across clusters, and it is a future research question whether an overall reliability measure makes sense and if not, how alternative reliability indices should be defined. Finally, in the current paper I assume that the sample cluster size is only a small fraction of the population cluster size when quantifying the sampling error of the between-level observed composite (Lüdtke et al., 2011). However, when the population cluster size is small (e.g., 25 students in a classroom) and the sample cluster size represents a large proportion of the population cluster size, the sampling error variance needs to be adjusted by a finite population correction factor (e.g., Lai, Kwok, Hsiao, & Cao, 2018). Future research is needed to discuss how reliability indices can incorporate such a correction factor for finite clusters.

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology. *American Psychologist*, *73*(1), 3–25. https://dx.doi.org/10.1037/amp0000191

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling*, *26*(1), 119–142. https://dx.doi.org/10.1080/10705511.2018.1511375

Brunner, M., & Süß, H. M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement*, *65*, 227–240. https://dx.doi.org/10.1177/0013164404268669

Conger, A. J. (1980). Maximally reliable composites for unidimensional measures. *Educational and Psychological Measurement*, *40*(2), 367–375. https://dx.doi.org/10.1177/001316448004000213

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://dx.doi.org/10.1007/BF02310555

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138. https://dx.doi.org/10.1037/1082-989X.12.2.121

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://dx.doi.org/10.1177/1948550617693063

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://dx.doi.org/10.1037/a0032138

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—Festschrift in honor of Karl Jöreskog* (pp. 195–216). Lincolnwood, IL:

Scientific Software International.

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 31–39. https://dx.doi.org/10.1080/10705511.2014.856694

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research*, *51*(6), 881–898. https://dx.doi.org/10.1080/00273171.2016.1228042

Knol, M. H., Dolan, C. V., Mellenbergh, G. J., & van der Maas, H. L. J. (2016). Measuring the quality of university lectures: Development and validation of the Instructional Skills Questionnaire (ISQ). *PLOS ONE*, *11*(2), e0149163. https://dx.doi.org/10.1371/journal.pone.0149163

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: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass. https://dx.doi.org/10.1177/001872679504800703

Kwok, O.-M., Underhill, A. T., Berry, J. W., Luo, W., Elliott, T. R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation psychology*, *53*(3), 370–386. https://dx.doi.org/10.1037/a0012765

Lai, M. H. C., Kwok, O.-m., Hsiao, Y.-Y., & Cao, Q. (2018). Finite population correction for two-level hierarchical linear models. *Psychological Methods*, *23*, 94–112. https://dx.doi.org/10.1037/met0000137

Liao, H., & Chuang, A. (2007). Transforming service employees and climate: A multilevel, multisource examination of transformational leadership in building long-term service relationships. *Journal of Applied Psychology*, *92*(4), 1006–1019. https://dx.doi.org/10.1037/0021-9010.92.4.1006

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*(4), 325–336. https://dx.doi.org/10.1177/001316445501500401

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. https://dx.doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203–229. https://dx.doi.org/10.1037/a0012869

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. https://dx.doi.org/10.1037/0022-0663.79.3.280

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124. https://dx.doi.org/10.1080/00461520.2012.670488

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://dx.doi.org/10.1037/1082-989X.10.3.259

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS Advanced 2015 international results in advanced mathematics and physics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from `https://timssandpirls.bc.edu/timss2015/advanced/`

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible

representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://dx.doi.org/10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo Confidence Intervals for Indirect Effects. *Communication Methods and Measures*, *6*(2), 77–98. https://dx.doi.org/10.1080/19312458.2012.679848

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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*(2), 173–184. https://dx.doi.org/10.1177/01466216970212006

Raykov, T., & du Toit, S. H. C. (2005). Estimation of reliability for multiple-component measuring instruments in hierarchical designs. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(4), 536–550. https://dx.doi.org/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: A Multidisciplinary Journal*, *13*(1), 130–141. https://dx.doi.org/10.1207/s15328007sem1301_7

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from `http://www.jstatsoft.org/v48/i02/`

Rush, J., & Hofer, S. M. (2014). Differences in within- and between-person factor structure of positive and negative affect: Analysis of two intensive measurement studies using multilevel structural equation modeling. *Psychological Assessment*, *26*(2), 462–473. https://dx.doi.org/10.1037/a0035666

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(4), 583–601. https://dx.doi.org/10.1080/10705510903203466

Stapleton, L. M., & Johnson, T. L. (2019). Models to examine the validity of cluster-level factor structure using individual-level data. *Advances in Methods and Practices in Psychological Science*, 251524591985503. https://dx.doi.org/10.1177/2515245919855039

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://dx.doi.org/10.3102/1076998616646200

Thompson, B. (2003). *Score Reliability: Contemporary thinking on reliability issues*. Thousand Oaks, CA: Sage.

Williams, T., Ferraro, D., Roey, S., Brenwald, S., Kastberg, D., Jocelyn, L., . . . Gonzales, P. (2009). *TIMSS 2007 U.S. technical report and user guide* (Tech. Rep. No. NCES 2009-012). U.S. Department of Education. Retrieved from https://nces.ed.gov/pubs2009/2009012.pdf

Footnotes

[1]On November 29, 2019, which was about five years following the publication of Geldhof et al. (2014), the paper has been cited 461 times according to Google Scholar.

[2]Also, Geldhof et al. (2014) noted that H had suboptimal performance and did not recommend its use. It is not clear, however, whether the performance would improve if the multilevel extensions of H takes into account the sampling variability of the cluster means for the between-level weighted composites.

[3]See Raudenbush and Bryk (2002, p. 46) for the reliability of cluster-level estimates for variables that are assumed to be measurement error-free.

[4]Stapleton et al. (2016) also discussed a more general model where the within-level variance of η^w can vary across clusters, but for simplicity and as a common practice, I assumed homogeneous variances of η^w across clusters in this paper.

[5]The items were scored on a 4-point scale, so MCFA for continuous variables could lead to biases in factor loadings and also in reliability of composite scores. I used continuous MCFA only so that results can be compared to Geldhof et al. (2014). Composite reliability for categorical items are beyond the scope of this paper.

Table 1
Interitem Covariances and Correlations of Attitude Toward Math Items

	1	2	3	4
		Within-Level		
AS4MAMOR	1.17	*.62*	*.61*	*.51*
AS4MAENJ	0.64	0.92	*.74*	*.59*
AS4MALIK	0.68	0.73	1.05	*.59*
AS4MABORr	0.59	0.61	0.66	1.18
		Between-Level		
AS4MAMOR	0.11	*.94*	*.89*	*.83*
AS4MAENJ	0.07	0.05	*.98*	*.95*
AS4MALIK	0.07	0.06	0.06	*.95*
AS4MABORr	0.06	0.05	0.05	0.05

Note. The covariances (correlations) were shown in the lower (upper) diagonal entries.

Table 2
Parameter Estimates of a Multilevel Confirmatory Factor Model for the Empirical Illustration

	Estimate	*SE*	LL	UL
λ_1^w	0.79	0.01	0.76	0.81
λ_2^w	0.82	0.01	0.80	0.84
λ_3^w	0.88	0.01	0.86	0.90
λ_4^w	0.75	0.01	0.73	0.77
θ_{11}^w	0.56	0.01	0.54	0.58
θ_{22}^w	0.24	0.01	0.23	0.25
θ_{33}^w	0.27	0.01	0.26	0.29
θ_{44}^w	0.62	0.01	0.60	0.64
ϕ_{11}^w	1.00	0.00	1.00	1.00
λ_1^b	0.79	0.01	0.76	0.81
λ_2^b	0.82	0.01	0.80	0.84
λ_3^b	0.88	0.01	0.86	0.90
λ_4^b	0.75	0.01	0.73	0.77
θ_{11}^b	0.03	0.00	0.02	0.04
θ_{22}^b	0.00	0.00	0.00	0.00
θ_{33}^b	0.00	0.00	−0.00	0.00
θ_{44}^b	0.01	0.00	−0.00	0.01
ϕ_{11}^b	0.08	0.01	0.06	0.10

Note. *SE* = Standard Error. LL, UL = lower and upper limits of 95% confidence intervals.

Table 3
Reliability Coefficients for Composite Scores in the Real Data Illustration, and Their Estimates for the Empirical Example

Type of Composites	Model	Reliability Coefficients	Equation in text	Est	95% Wald CI	95% MC CI	95% Bayesian CrI
Overall[a]	Individual	ω^{2l}	(13)	.87	[.86, .87]	[.86, .87]	[.86, .87]
	—	α^{2l}	(23)	.87	[.86, .87]	[.86, .87]	[.86, .87]
Between-Level[b]	Configural	ω^{b}	(17)	.62	[.56, .68]	[.55, .68]	[.56, .69]
	—	α^{b}	(24)	.65	[.59, .71]	[.58, .70]	[.60, .72]
Within-Cluster	Individual	ω^{w}	(8), (15)	.86	[.86, .87]	[.86, .87]	[.86, .87]
	Within	ω^{w}	(21)	.86	[.86, .87]	[.85, .87]	[.86, .87]
	—	α^{w}	(25)	.86	[.85, .87]	[.85, .86]	[.85, .87]

Note. Model specification is only relevant for ω but not for α. Est = Reliability estimate. MC = Monte Carlo. CI = Confidence Interval. CrI = Credible Interval.

[a]When measuring individual constructs, it is recommended to report all three of ω^{2l}, ω^{b}, and ω^{w} (or all three of α^{2l}, α^{b}, and α^{w}).

[b]See the Appendix for an example of a shared construct.

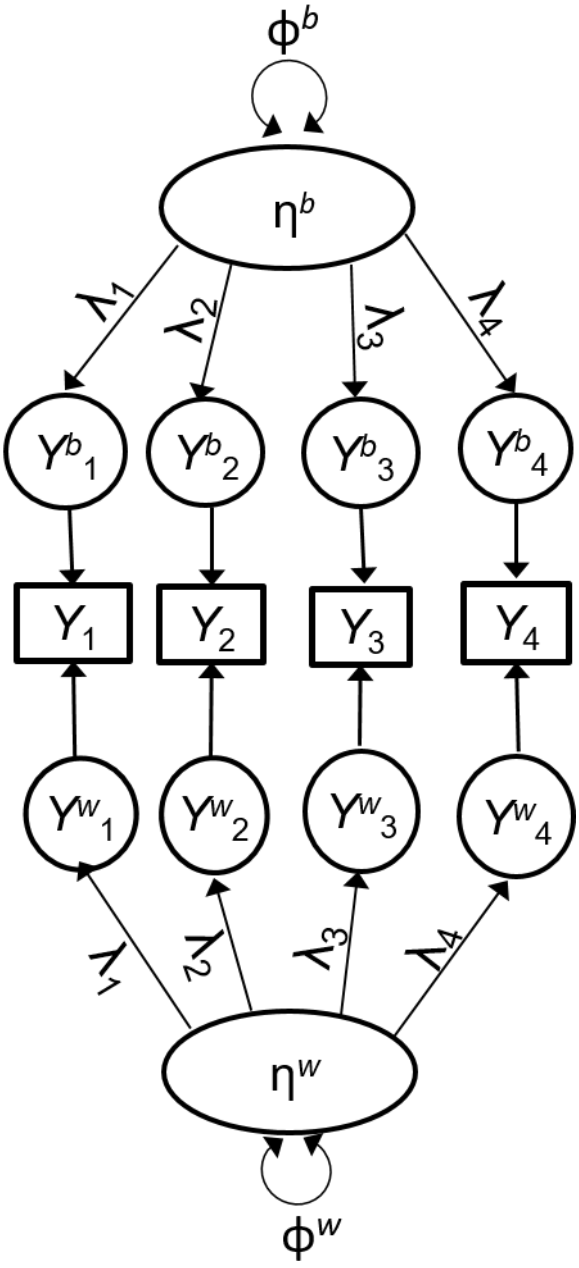

Figure 1. Path diagram depicting a multilevel factor model for an individual-level construct. This is a modified re-creation of Figure 5 in Stapleton et al. (2016).

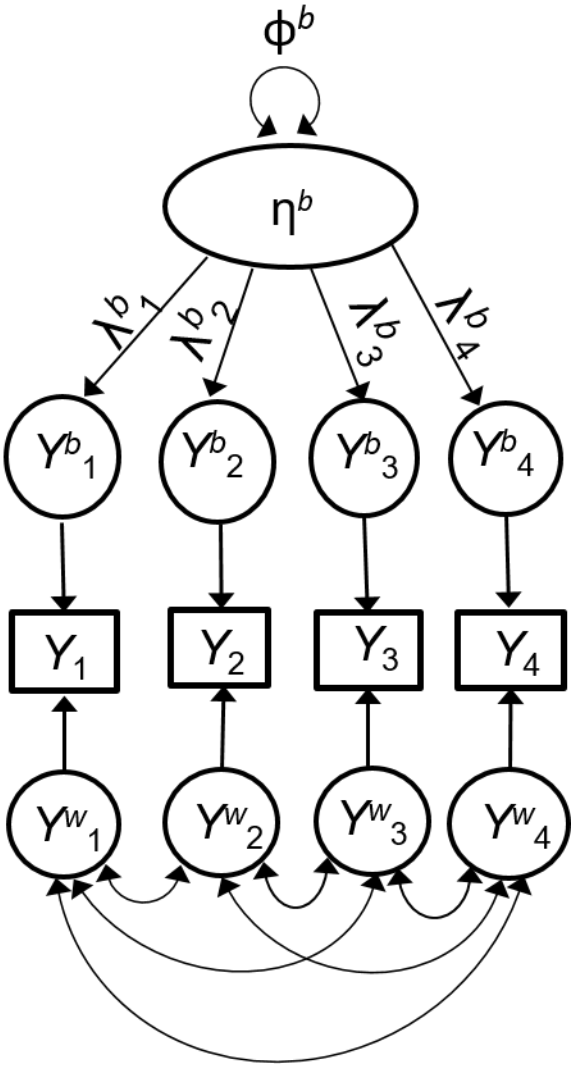

Figure 2. Path diagram depicting a multilevel factor model for a shared construct. This is a modified re-creation of Figure 4 in Stapleton et al. (2016).

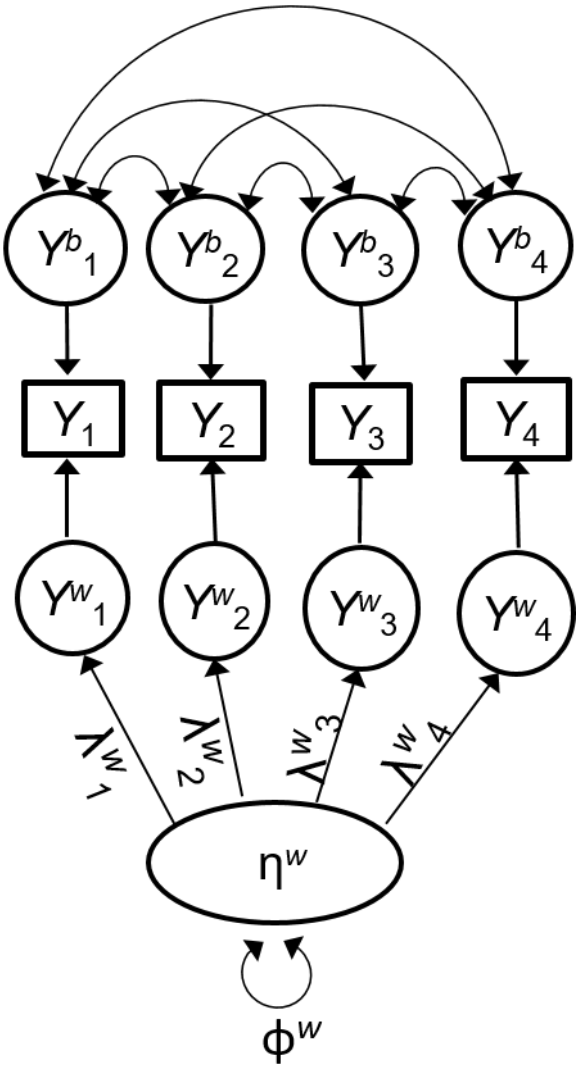

Figure 3. Path diagram depicting a multilevel factor model for a within-cluster construct. This is a modified re-creation of Figure 3 in Stapleton et al. (2016).

Appendix

Reliability of a Shared Construct (With the Potential Presence of a Configural Construct)

To illustrate the computation and issues for the reliability of a between-level composite measuring a shared construct, I revisited the empirical example in (Stapleton & Johnson, 2019) based on the TIMSS 2015 data (Mullis, Martin, Foy, & Hooper, 2016) on six student-level items that tapped into the teacher-level shared construct Engaging Teaching. After listwise deletion, the data contained 2,891 students from 240 classrooms. The six items asked whether students agreed with the following statements (1-*agree a lot* to 4-*disagree a lot*):

1. The teacher clearly communicates the purpose of each mathematics lesson (`MSBM18A`).
2. I know what my teacher expects me to do (`MSBM18B`).
3. My teacher is easy to understand (`MSBM18C`).
4. My teacher links new content to what I already know (`MSBM18H`).
5. My teacher is good at explaining advanced mathematics (`MSBM18I`).
6. My teacher uses a variety of teaching methods, tasks, and activities to help us learn (`MSBM18M`).

The ICCs of the items were between .10 to .21. Other parameter estimates can be found in Stapleton and Johnson.

If one is only interested in the between-level shared construct, and would like to use the between-level composite to measure it, one should use a saturated within-level model (Figure 2). Using equation (20), the estimated $\omega^b = .719$, 95% Wald CI [.668, .771], which was very different from $\tilde{\omega}^b = .976$ reported in Stapleton and Johnson (2019).

However, Stapleton and Johnson (2019) also hypothesized that there was an additional individual-level construct, Acquiescence, as a source of common variance at the individual level (see Figure A1). If Acquiescence had between-level variance (i.e., ICC > 0), then the common variance at the between level would not be unidimensional. As explained in Stapleton and Johnson, given that Engaging Teaching, denoted as η^s, and the between-level component of Acquiescence, denoted as η^b, had the same set of indicators, their respective variances, ϕ^s and ϕ^b,

were not estimable. One solution that Stapleton and Johnson proposed was to directly fix the ICC of Acquiescence to a chosen value based on researchers' knowledge.

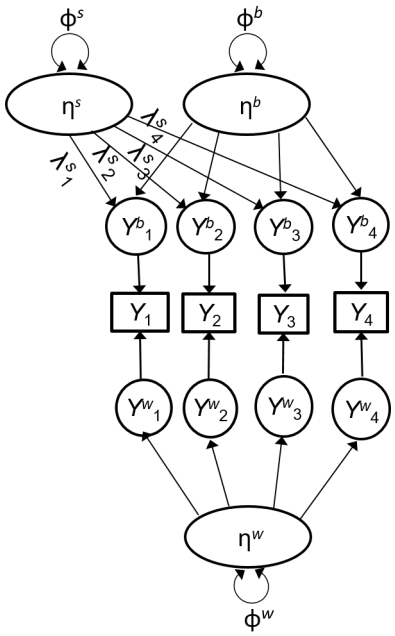

Figure A1. Path diagram depicting a multilevel factor model for a simultaneous shared and in individual-level construct. This is a modified re-creation of Figure 6 in Stapleton et al. (2016).

The fact that the between-level composite score might measure more than one construct does not change the definition of the definition of reliability, at least in classical test theory, because reliability is simply the consistency of the observed composite scores, even when the those scores measure also something irrelevant to the target construct. In other words, ω^b and α^b correctly measures the consistency of the between-level composite, but the composite itself may measure some stable cluster-level characteristics (e.g., classroom-level aggregates of Acquiescence) other than the target shared construct (e.g., Engaging Teaching). The presence of an additional configural construct does have an impact on the validity of the observed between-level composite scores, and in such a situation researchers may be more interested in the proportion of observed variance attributable to the target shared construct. Let λ^s_k be the factor loading of the kth item on the shared construct, then an index quantifying the proportion of

variance attributable to the shared construct is

$$\omega^{b(s)} = \frac{(\sum_{k=1}^{p} \lambda_k^s)^2 \phi^s}{(\sum_{k=1}^{p} \lambda_k^s)^2 \phi^s + (\sum_{k=1}^{p} \lambda_k)^2 (\phi^b + \phi^w/\tilde{n}) + \mathbf{1}'\boldsymbol{\Theta}^b\mathbf{1} + \mathbf{1}'\boldsymbol{\Theta}^w\mathbf{1}/\tilde{n}}, \tag{A1}$$

where λ_k is the factor loading of the kth item on the individual construct (and is constant across the configural and the within-level component). Note that $\omega^{b(s)}$ is conceptually the same as what Stapleton and Johnson (2019) referred to as *construct reliability* (see also Brunner & Süß, 2005).

Based on one of the scenario demonstrated in Stapleton and Johnson (2019) that assumed an ICC of .05 for Acquiescence, $\omega^{b(s)}$ was estimated to be .607, 95% Wald CI [.534, .679]. However, if the ICC of Acquiescence was assumed to be .10, $\omega^{b(s)}$ would be .476, 95% Wald CI [.378, .575], as less variance was due to the shared construct. Therefore, the reliability of the between-level composite, ω^b, is an upper bound for $\omega^{b(s)}$. The R and Mplus codes for computing ω^b and $\omega^{b(s)}$ for this example can be found in the supplemental material.