Article

Journal of Educational and Behavioral Statistics
2016, Vol. 41, No. 5, pp. 481–520
DOI: 10.3102/1076998616646200

Construct Meaning in Multilevel Settings

Laura M. Stapleton
Ji Seung Yang
Gregory R. Hancock
University of Maryland

We present types of constructs, individual- and cluster-level, and their confirmatory factor analytic validation models when data are from individuals nested within clusters. When a construct is theoretically individual level, spurious construct-irrelevant dependency in the data may appear to signal cluster-level dependency; in such cases, however, and consistent with theory, a single-level analysis with a correction for dependency may be appropriate. Regarding cluster-level constructs, we discuss two types—shared and configural—and present appropriate validation models. Illustrative validation analyses with individual, shared, and configural constructs are provided using empirical data as well as simple simulations demonstrating the spurious effects that can occur with nested data. The article concludes with future directions to be examined in construct validation in multilevel settings.

Keywords: *multilevel; validity; confirmatory factor analysis*

In social science research, latent constructs of interest are typically validated using not only theoretical arguments but also empirical data from sets of items intended to measure those constructs. When data are collected in multilevel settings (e.g., students within schools or children within families), a construct of interest might even exist at multiple levels. In this article, we consider how researchers can approach construct meaning and construct validation when working with data that are nested. In Section 1, we briefly review the confirmatory factor analysis (CFA) approach to structural validation of a construct hypothesized to underlie multiple item responses and discuss the extension of the single-level approach to a simple multilevel CFA (MCFA) when data are nested. In Section 2, we try to bring clarity to the murky conceptual landscape that exists when considering measurement models at both the individual and the cluster levels, emphasizing distinctions between constructs at different levels and, importantly, between different types of constructs at the cluster level. In this section, we define five distinct models that might be posited using the same data, depending on conceptual considerations. In Section 3, we present examples of construct structure validation processes in several contexts using data from the National Center for Education Statistics, while in Section 4, we present simulation demonstrations of contexts

where, with an individual-level construct, spurious intraclass correlations (ICCs) and spurious cluster-level covariance can result. Finally, we suggest practical modeling guidelines and conclude with future extensions that may be considered.

1. Validation Evidence for the Structure of a Measure

Measurement of a person's motivation, of a child's knowledge level, or of a teacher's degree of job satisfaction is typically accomplished via the combination of responses to several questionnaire or test items. Just because multiple item responses are used, however, does not assure that any single aggregate measure of those responses has adequate validity. Messick (1989, 1995) argued that a thorough assessment of the use of a measure should attend to, and provide evidence for, many aspects of construct validity; central to the current article is the *structural* aspect, which addresses the fidelity of any scoring structure (moving from items to an aggregate score), including the aggregation mechanism, any scoring criteria and rubrics, and the application of those criteria and rubrics.

1.1. Single-Level CFA and Reliability

Currently, the use of CFA is ubiquitous when documenting the dimensionality and reliability of scores on an instrument designed to measure psychological processes or states (Brown, 2015). As a simple example, suppose a researcher develops a set of 4 items intended to tap some construct (ξ) representing a dimension of a person's opinion or belief. The researcher hopes to use a combination of these items (e.g., a sum, average, or weighted composite) to represent the construct in future research. In this case, a CFA process can be used to determine whether there is empirical support for the hypothesized construct's measurement. A model for this construct, ignoring the mean structure for simplicity, may be shown as in Figure 1, with the accompanying measurement equation:

$$\mathbf{x} = \boldsymbol{\lambda}\xi + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{x}$ is a $p \times 1$ vector of observed responses from a respondent on p items, $\boldsymbol{\lambda}$ is a $p \times 1$ vector of loadings relating the items to the underlying construct ξ, and $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of error terms for the observed variables with an assumed multivariate normal covariance matrix $\boldsymbol{\Theta}$ (which is diagonal in this example but is not generally required to be so). Although we have presented the measurement structure in terms of a single construct here and throughout this article, the model extends to multiple constructs.

If such a model is found to be plausible, given the observed covariance matrix for the item responses in a sample data set, then some support for

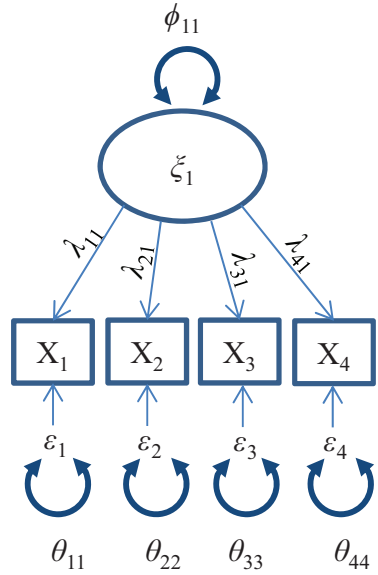

FIGURE 1. *Single-level factor model.*

the hypothesized structure of the measure is gained. Reliability of latent construct score estimates can then be estimated by a range of measures, including those that use parameter estimates from the CFA. One measure, utilized in this article, is *composite reliability*, ω (McDonald, 1970, 1978, 1999; see also Raykov, 1997), which is defined as follows: where i indexes the item, ϕ refers to the factor variance, and θ represents the residual item variance.

$$\omega = \frac{\left(\sum_{i}^{p} \lambda_i\right)^2 \phi}{\left(\sum_{i}^{p} \lambda_i\right)^2 \phi + \sum_{i=1}^{p} \theta_i}. \tag{2}$$

While the internal consistency coefficient α (Cronbach, 1951) is often used as a measure of scale reliability, it assumes that all items load on a single underlying construct and are τ equivalent (i.e., all items are equally correlated with the underlying construct). Alternatively, ω assumes only that there is a congeneric scale and allows items to demonstrate variability in strength of relation with the underlying construct. While maximal reliability has also been proposed (see Hancock & Mueller, 2001), its performance in multilevel models has been questioned (Geldof, Preacher, & Zyphur, 2014) and is not addressed in this article.

1.2. Multilevel CFA and Reliability

A complication arises, however, when data are collected within nested settings, such as students within schools and children within families. The possibility for structural validation can occur at both levels as well as at an aggregate overall level; indeed, multilevel researchers have recently suggested conducting the validation at both the within-cluster and between-cluster levels (Geldhof, Preacher, & Zyphur, 2014; Forer & Zumbo, 2011; Zyphur, Kaplan, & Christian, 2008). In this MCFA approach, each observed variable is parsed into within and between components.

$$\mathbf{x} = \boldsymbol{\eta}_W + \boldsymbol{\eta}_B, \tag{3}$$

where $\boldsymbol{\eta}_W$ represents a $p \times 1$ vector of cluster mean–centered deviations (or within-cluster processes) and $\boldsymbol{\eta}_B$ represents a $p \times 1$ vector of latent cluster means (or between-cluster processes). The measurement model is then hypothesized on two covariance matrices, one for within-cluster (cluster mean centered) variability and the other for between-cluster variability (conceptually related to covariances of the cluster means; see Muthén, 1991):

$$\boldsymbol{\eta}_W = \boldsymbol{\lambda}_W \xi_W + \boldsymbol{\varepsilon}_W \quad \text{and} \tag{4}$$

$$\boldsymbol{\eta}_B = \boldsymbol{\lambda}_B \xi_B + \boldsymbol{\varepsilon}_B. \tag{5}$$

The combined measurement model, linking the observed variables to the underlying latent factors, thus becomes:

$$\mathbf{x} = \boldsymbol{\lambda}_W \boldsymbol{\xi}_W + \boldsymbol{\varepsilon}_{\mathrm{W}} + \boldsymbol{\lambda}_B \boldsymbol{\xi}_B + \boldsymbol{\varepsilon}_B, \tag{6}$$

where each component of the single-level CFA model is now represented at both the within-cluster and between-cluster levels, as shown in Figure 2 for our context of a unidimensional latent structure at each level. Dashed circles are used to represent the components that have been separated into within-cluster and latent cluster-level processes. The within-cluster-level structure is shown in the bottom half of Figure 2 and the between-cluster structure is shown at the top, with subscripts of *W* and *B*, respectively.

Literature exists evaluating the statistical equivalence of these models with hierarchical linear models (Li, Duncan, Harmer, Acock, & Stoolmiller, 1998; Mehta & Neale, 2005), and reliability measures for latent scores from these models have been proposed (Raykov, 2009; Raykov & Marcoulides, 2006; Raykov & Penev, 2010). Raykov's procedures, however, assume a single reliability value aggregated over both the within-level and between-level structure and thus assume an individual construct with no cluster-level construct that influences item responses above and beyond differences in the latent means for individuals across clusters. We refer to this assumed condition as the existence of only a *configural* cluster-level factor and expand on its

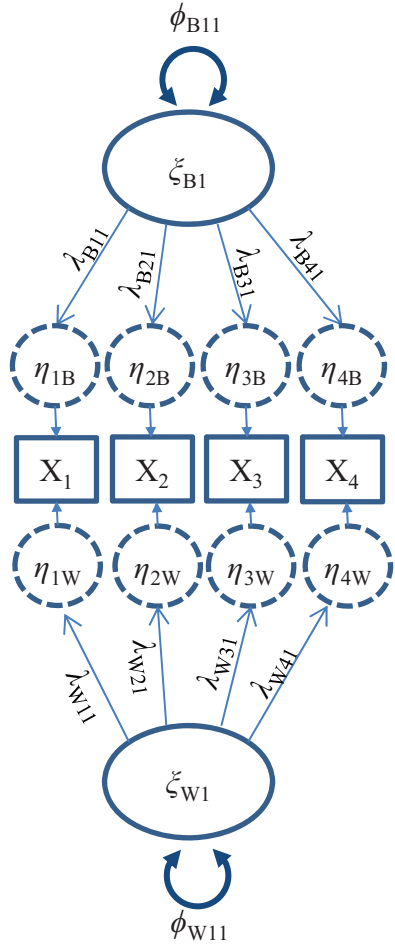

FIGURE 2. *Two-level factor model.*
Note: Each η *component also has a residual with variance* θ*, not shown for simplicity.*

definition in Section 2. Geldhof, Preacher, and Zyphur (2014) have proposed to extend estimation of ω to the within and between portions of the model and thus any set of items might have two reliability estimates: within-cluster and between-cluster.

In nested data settings, with the measurement of cluster-level constructs, a variety of measures are used to evaluate whether item responses show some degree of clustering as would be expected for a construct at the cluster level. One such oft-used measure is the ICC, also referred to as ICC(1) (Shrout &

Fleiss, 1979). ICC(1) for a single manifest item is defined as:

$$\text{ICC}(1) = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \sigma^2_{\eta_W}}, \tag{7}$$

where $\sigma^2_{\eta_B}$ represents the variability of the cluster-level component and $\sigma^2_{\eta_W}$ represents the variability of the within-cluster component and is thus interpreted as the proportion of variance in an observed variable that is found at the cluster level. Shrout and Fleiss presented a second measure of clustering, referred to as ICC(2) and used in multilevel models as a measure of reliability of cluster components (Raudenbush & Bryk, 2002), estimated as:

$$\text{ICC}(2) = \frac{\sigma^2_{\eta_B}}{\sigma^2_{\eta_B} + \frac{\sigma^2_{\eta_W}}{n.}}, \tag{8}$$

where *n.* is the average cluster size for an estimate of average reliability over all clusters or where *n.* is n_j to obtain a reliability estimate for a given cluster *j*. The ICC(2) estimate can thus be used to determine the number of measures needed from a given cluster to result in a sufficient level of reliability (Raudenbush & Bryk, 2002). Additionally, ICC(2) is mathematically equivalent to a special case of the generalizability coefficient for group means when persons are nested within clusters and crossed with fixed items (see Kane & Brennan, 1977, for more details). The richer discussion on reliability of group mean scores in the context of norm- and criterion-referenced interpretation is available in the framework of generalizability theory where sources of variance in observed scores are specified and portioned to conceptualize reliability and are available in Cronbach, Gleser, Nanda, and Rajaratnam (1972), O'Brien (1990), and Brennan (1995). We do not include the analogical discussion on reliability with respect to latent variables in MCFA models in this article, as our focus is on structural validity; however, reliability of latent group means deserves further extensive development and discussion in future work.

Somewhat absent in the educational research literature regarding these MCFA measurement approaches, however, is a discussion of whether a multilevel approach to the validation exercise is even conceptually appropriate. In the next section, we propose five CFA models that might be utilized with nested data; as will be explained, a key distinction among these models is whether it is hypothesized that the construct of interest exists at an individual level only or is of interest at the cluster level as well.

2. Conceptual Issues in Structural Validation in a Multilevel Context

We argue that three central issues must be addressed when obtaining evidence of measurement structure under nested data settings: (1) the ultimate unit (or

level) of interest in the measurement process (individual or cluster levels); and if interested in individual-level constructs (2) the causes of any homogeneity in item responses within clusters and (3) the intended uses of scores. These issues are addressed within the two large subsections that follow, one focused on individual-level constructs and one on cluster-level constructs. Overall, we propose considering five different types of models, as listed in Table 1, when considering construct validation with nested data. Across the top of the columns are the model names and within each column we display the level-specific construct that is (or is not) modeled and interpreted, possible use of the construct, and related treatments in the literature. Each of these model types is described in detail in the subsections that follow.

2.1. Individual as Unit of Interest

At the individual level, it is expected that the hypothesized construct is relevant to the individual responding, and therefore scores on the measure should reflect individual variability. Models 1, 2, 4, and 5 in Table 1 include such a construct. There may be no assumption that a cluster-level true score (what we will refer to as a *shared* cluster-level construct) exists; some researchers, however, have implied that whenever manifest item scores exhibit clustering, a true score at the cluster level is responsible for part of the score (Bliese, 2000; Forer & Zumbo, 2011; Geldhof et al., 2014). Geldhof et al. stated that for construct validation, "level-specific reliability estimates ... are generally preferable to single-level estimates whenever ICCs are nontrivial (i.e., $\geq$.05)" (p. 89). Similarly, Bliese suggested that "when an individual measure ... has an ICC(1) value larger than zero, this indicates that an aggregate variable will partially reflect common environmental factors" (p. 374). Forer and Zumbo also stated that the ICC(1) "represents the proportion of individual variance that is influenced by, or depends on group membership" (p. 241). In reading those statements, an applied researcher might infer that a true cluster construct must exist that causes or influences the environment. However, such an impression would imply that moving an individual from one cluster to another would result in a modification of that individual's item responses due to a difference in true cluster constructs. We argue, however, that researchers should consider the cause of the homogeneity within clusters prior to settling on an MCFA approach to validate measurement structure.

Suppose a researcher was evaluating a measure of lactose intolerance for use in a school-aged population. A researcher might have developed five questionnaire items hypothesized to tap the latent lactose intolerance construct and, for validation purposes, administered the items to students who are nested within schools. The items inquire about the degree of severity, after consumption of products containing lactose, of a set of physical reactions: diarrhea, abdominal cramping, vomiting, audible bowel sounds, and flatulence or gas (Casellas,

TABLE 1.
Five Possible Multilevel CFA Models Based on Individual-Level Measures

Considerations for the Construct	Model 1 Individual (Single-Level CFA Model)	Model 2 Within Cluster	Model 3 Shared	Model 4 Configural	Model 5 Shared and Configural
Level at which construct is of interest	Aggregate	Within cluster	Between cluster	Both within and between cluster	Both within and between cluster
Interpretation at between-cluster level	Covariance is combined across levels; clustering effect is assumed to be spurious	Construct is not of interest; saturated model	Construct represents a characteristic of the cluster	Construct represents the aggregate of the disparate characteristics of individuals within a cluster; cross-level invariance constraints required	Covariance includes both a true cluster-level construct and one that is an aggregate of individual characteristics (for which cross-level invariance constraints are required)
Interpretation at within-cluster level	Covariance is combined across levels; clustering effect is assumed to be spurious	Model imposed within cluster; represents individual standing within cluster	Covariation should not exist; responses only reflect cluster-level construct	Model imposed within cluster; with cross-level invariance constraints	Model imposed within-cluster; with cross-level invariance constraints

(continued)

TABLE 1. *(continued)*

Considerations for the Construct	Model 1 Individual (Single-Level CFA Model)	Model 2 Within Cluster	Model 3 Shared	Model 4 Configural	Model 5 Shared and Configural
Future use of measure	To compare individuals across a broad population	To compare individuals within clusters	As a cluster characteristic for organizational climate research	Both for within-cluster comparisons of individuals and for cluster-level contextual effect studies	For within-cluster comparisons of individuals as well as for cluster-level climate and contextual effect studies
Alternate names in literature			Climate (Marsh et al., 2012); reflective (Lüdtke et al., 2011)	Contextual (Marsh et al.); formative (Lüdtke et al.)	Not discussed in the literature

Note. CFA = confirmatory factor analysis.

Varela, Aparici, Casaus, & Rodriguez, 2009). We contend that lactose intolerance is a completely individual measure, not able to be influenced by a school or environmental effect. It is very likely, however, that responses to these items will display a nonnegligible ICC. The reason is that schools will vary in terms of their membership according to ethnic origin of the individuals. Medical research has confirmed that lactose maldigestion differs by subpopulation depending on geographic origin, sometimes greatly so, with those of European origin having the lowest severity of the condition (Scrimshaw & Murray, 1988). Because schools differ in terms of the proportion of students by region of origin, levels of responses to each of the five symptom items will differ as well, resulting in an item ICC(1) above zero. We consider this positive ICC(1) a *spurious ICC*, a function of a selection process. When hypothesizing a latent construct, a researcher should bring a clear theoretical argument regarding whether a measure demonstrates a positive ICC due to a true cluster construct or whether the positive ICC represents a spurious relation. As demonstrated in Section 4, the observed statistics in a single study do not provide information to the researcher regarding which of these situations is correct. Thus, in the absence of a study moving individuals across clusters to examine whether their levels on a response variable change or the availability of auxiliary data that may shed light on possible existence of informative subpopulations within clusters, theory is absolutely crucial.

If a researcher believes that any cluster dependency is spurious, then a design-based approach (Stapleton, 2013) to validation of a measure's structure is appropriate. A design-based approach allows an analyst to adjust the standard error estimates and model fit indices, given the dependency of item responses of individuals within clusters. The model parameter estimates will provide a single-level, aggregate of the relation between item responses and constructs. While this approach may obscure possible cross-level noninvariance (Zyphur et al., 2008), its use may be appropriate in conditions such as those demonstrated in Section 4. Two design-based estimation approaches for CFA that are currently available to the researcher in many software programs are linearization (Asparouhov & Muthén, 2005; Stapleton, 2006) and replication (Asparouhov & Muthén, 2010; Stapleton, 2008). These methods are briefly explained in Section 1 of the supplement materials, available in the online version of the journal.

An issue that must be addressed when working with nested data and wishing to measure a construct at the individual level only is to consider the future use of the measure. This issue is related to determining the (sub)population for which it is important to validate the measure. If a researcher was planning to use the measure across a broad population in the future (that may or may not be nested within clusters) and intended to use the measure in a single-level model, then the model shown in the first column of Table 1 with a design-based estimation approach to validation is appropriate. If, however, the researcher intended to use the measure within a cluster, to compare individuals and their relative position within only a specific cluster, then a within-cluster approach to validation would

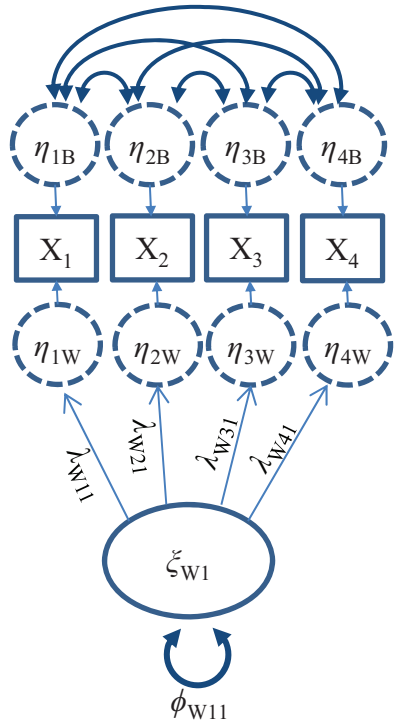

FIGURE 3. *Within-cluster construct model.*
Note: Each within-cluster η *component also has a residual with variance* θ*, not shown for simplicity.*

provide within-cluster reliability coefficients that are more suitable to the future planned utilization of the measure. This approach, referred to in Table 1 as Model 2 (within cluster), can be conducted in one of the two ways: analysis of cluster-centered data or an MCFA. In an analysis of cluster-centered data, a single-level CFA can be conducted based on a pooled within-cluster covariance matrix (Hox, 2002). This type of analysis is very straightforward and it would restrict the interpretation to the structure of responses relative to others in the same cluster. To conduct an MCFA, the model shown in Figure 3 could be estimated. This model does not assume existence of a cluster-level construct but does allow for cluster-level variability on each measure with a saturated model of the covariances among them. The within-cluster covariation is used to test the plausibility of a within-cluster construct that may be used in the future to compare individuals who share a cluster or to identify relations among constructs within a cluster. For example, if future interest were in modeling whether a school principal might use a student's lactose intolerance scores to predict his family's interest in

participating in school lunch programs (relative to other students in the school), then this model would be appropriate for validation of such scores.

2.2. Cluster as Unit of Interest

Item responses from individuals can also be used to measure constructs at the cluster level and we will discuss two types of constructs: one that is a characteristic of the cluster itself and one that is just a reflection of the construct at the individual level. The nature of a latent construct at the cluster level can be perplexing, and the importance of considering the interpretation of construct meaning at each level was first brought forward by Cronbach (1976) and conceptually developed in organizational research (see Chan, 1998; Kozlowski & Klein, 2000) but has not been squarely addressed by many recent MCFA-applied analyses (see, as e.g., Dedrick & Greenbaum, 2011; Klangphahol, Traiwichitkhum, & Kanchanawasi, 2010). Marsh and colleagues (2012) highlighted this oversight for many multilevel modelers who conduct contextual research using structural equation models and suggested the differences between predictors that represent *climate* constructs and *contextual* constructs at the cluster level, where climate refers to a shared experience and context refers to the aggregation of disparate individual responses. These constructs have also been differentiated by the terms *reflective* and *formative*, respectively (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). Kozlowski and Klein (2000) also differentiated cluster-level constructs into two types, essentially with the same meaning as that provided by Marsh et al. and labeled them as shared and configural, terms that we will use in this article and are shown in our typology in Table 1 as Models 3 and 4, respectively. The critical distinction between the two types of constructs has been well formulated in organizational psychology and we detail the differences below.

2.2.1. Shared cluster constructs. Shared constructs can be measured using individual-level item responses that are intended to measure a characteristic of the cluster. For a cluster-level shared construct, one would expect individuals within the cluster to respond in a similar way if the measurement tool provides valid scores. For a truly shared construct, the measures would be isomorphic across individuals in the cluster; given this isomorphism, then, the only measure of interest at the cluster level would be the mean response of the individuals in the cluster. As an example, to measure instructional quality, a characteristic of the classroom and not of the individual student, responses to items from students in the same classroom should be highly correlated; in fact, they should be seen as interchangeable. A shared construct could therefore be validated by imposing Model 3 in Table 1 as shown in Figure 4. Any variability and covariation of responses at the within-cluster level are not of interest in this model; in fact, there should be minimal variability found at the within-cluster level for a truly shared construct. Bliese (2000) sets a very high criterion for considering a construct to

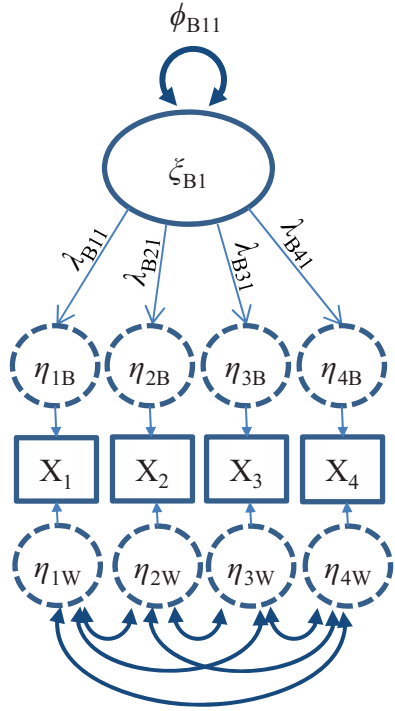

FIGURE 4. *Shared cluster construct model.*
Note: Each between-cluster η *component also has a residual with variance* θ_B*, not shown for simplicity.*

be a shared one, stating that "in the absence of within-group agreement, one's measurement model would be unsupported" (p. 367). Assuming adequate model fit,[1] the researcher could then assess reliability using a between-cluster measure of composite reliability as suggested by Geldhof et al. (2014), using Equation 2 but using only the cluster-level loading estimates and cluster-level factor variance and residual variance estimates.

For *manifest* variables that are hypothesized to be shared, it has been suggested that ICC(2) values (see Equation 8) of at least .7 represent acceptable levels of reliability of a measured shared construct, while ICC(2) values between .5 and .7 represent marginal reliability and values below .5 would be considered poor (Klein et al., 2000). The ICC(2) has not been extended to reliability estimation for a latent construct and would require that a construct exists at both the within-cluster and between-cluster levels (which is not the case for a shared construct model). However, we believe that a reasonable step in this shared construct modeling process would be to examine ICC(2) values for each item in a proposed latent shared measure to report along with the between-cluster composite reliability.

It should be noted that questionnaire item wording plays a crucial role in measurement in multilevel contexts. Item wording can cloud the differentiation of a construct as one that is individual or a shared cluster-level construct. When measuring a shared construct, without proper item stem wording, a researcher may obtain item responses that reflect both the shared and individual characteristics. For example, a question that posits "This instructor presents material in ways that keep it interesting" would likely elicit responses reflecting instructor qualities. A question that posits "I find the class meetings interesting" would likely reflect both a characteristic of the cluster (the teacher's instructional ability) and the rater's own intrinsic interest in the class topic. In the instrument development process, careful attention should be paid to item wording for the measurement of constructs at specific levels (Marsh et al., 2012).

In summary, a shared construct is measured when a researcher is interested in the level of a cluster characteristic, using individuals within clusters as the information source. If positing a shared cluster construct based on individual responses in a nested data setting, a researcher might provide the ICC(1) and ICC(2) estimates for each manifest response variable, evidence of support for the hypothesized relation between manifest variables and the latent construct at the cluster level via model fit information, and an estimate of the between-cluster composite reliability as defined by Geldhof et al. (2014). Furthermore, strong theoretical rationale should be provided if the CFA model also includes any construct modeled at the within-cluster level, in which case the researcher should consider the simultaneous shared and configural model (Model 5 in Table 1) described in Section 2.2.3.

2.2.2. Configural cluster constructs. Configural constructs, shown as Model 4 in Table 1, are cluster aggregates of individual constructs (e.g., average or dispersion of internalizing behavior of children within a family, motivation levels of children within a school). It is not expected that individuals within a cluster respond in the same way to the item measures, and their responses are not interchangeable across individuals. Marsh et al. (2012) proposed that such a model for these configural constructs could be measured as shown in the model first displayed in Figure 2, where the focus in this model is on the mean of the configural latent construct at the cluster level. Marsh et al. also suggested that if a measure has an ICC(1) value of zero, suggesting no variability in the mean item response across clusters, then there is little reason to continue with examining the configural cluster-level construct (p. 115). We argue, however, that it may be of interest to document the use of a measure to represent the variability in individual hypothetical latent scores or patterns in those individual latent scores across clusters. Marsh et al.'s advice ignores the idea that dispersion in the construct may differ across clusters and represent an important cluster characteristic. In fact, Kozlowski and Klein (2000) cautioned *not* to use the cluster mean alone to represent these configural constructs. Meade and Eby (2007) proposed what they

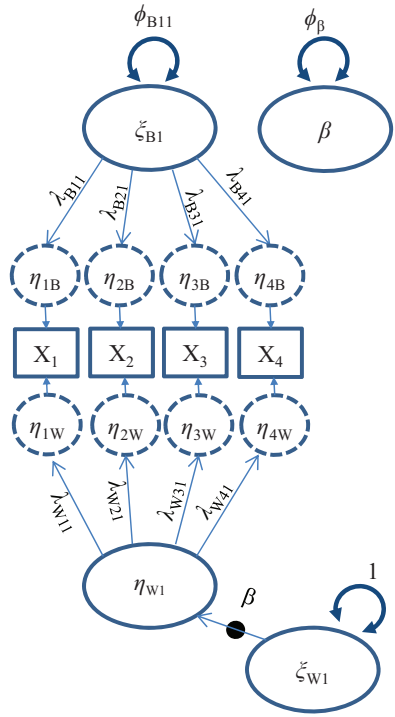

FIGURE 5. *Configural cluster construct model with required cross-level constraints* ($\lambda_{W11} = \lambda_{B11}$, $\lambda_{W21} = \lambda_{B21}$, $\lambda_{W31} = \lambda_{B31}$, $\lambda_{W41} = \lambda_{B41}$) *and with a measure of within-cluster dispersion* (β).
Note: Each item η *component also has residual with variance* θ_B *or* θ_W*, not shown for simplicity.*

termed a *dispersion model* approach when undertaking multilevel construct validation but calculated manifest measures of group agreement or heterogeneity outside of the CFA model. We propose, instead, to impose the model shown in Figure 5. In this model, the variance of the within-cluster latent variable is modeled as cluster-specific, accomplished by hypothesizing a random slope between the factor of interest and a phantom factor (ξ_{W1}) with unit variance and estimated utilizing Bayesian estimation (Asparouhov & Muthén, 2012; Levy & Choi, 2013). The model-implied variance of the within-cluster factor (η_W) is simply β_j^2 for the jth cluster, and the cluster-specific standard deviation, β_j, appears as a latent factor at the cluster level. Of specific interest in this model is whether the variance of the cluster-specific standard deviations, ϕ_β, differs from zero, indicating that clusters differ in the dispersion of the individuals within the cluster. This proposed dispersion model is but one option for modeling

cluster-specific differences in the distribution of the individual-level construct; alternative methods might be developed to hypothesize other types of pattern differences. This process of configural construct dispersion validation may be important if a researcher wants to develop a measure that is sensitive enough that it can be used to evaluate change in dispersion (e.g., if an intervention is intended to change a group dynamic and results in greater cohesiveness of thought or behavior).

With this modeling approach, there is an assumption that the cluster-level factor, $\xi_{\beta 1}$, is not a cluster shared construct but merely reflects the cluster aggregate of the individual construct at Level 1 (e.g., the average lactose intolerance level of children in a school). Therefore, the vector of factor loadings should be constrained across levels for each of the p observed variables ($\boldsymbol{\lambda}_W = \boldsymbol{\lambda}_B$). When a construct is at the individual level but differences in the mean construct exist across clusters because of spurious clustering (as was discussed in Section 2), there should be no difference in the unstandardized loadings across levels as the between-cluster relations just reflect the within-cluster relations, which has been referred to as *cross-level measurement invariance* (Zyphur et al., 2008). In the situation with spurious ICCs (such as our hypothetical lactose intolerance example), Models 1, 2, and 4 are equally appropriate, depending on whether the analyst desires an individual measure with broad applicability, a within-school interpretation of the individual measure, or a measure of the aggregate lactose intolerance (mean or dispersion) of students in the school, respectively.

In summary, a configural construct is defined as an aggregate of the measurements of individuals who comprise the cluster. Measures of interest may include dispersion of the construct as well as the mean level of the construct; the cluster itself is not viewed as the source or reason for variability of an individual construct and therefore between-cluster loadings are fixed to be the same as within-cluster loadings. Validation evidence might include the fit of the model as well as its ability to capture variability in dispersion across clusters.

2.2.3. Simultaneous shared and configural cluster constructs. It is possible that more than one construct is required to adequately model the cluster-level covariation. For example, suppose that teachers have provided ratings of their students' motivation using 4 items for each child, it is possible that some teachers tend to rate more positively as compared to others. The obtained data would then contain both sources of covariation at the cluster level (variation due to the fact that in some classes, students are truly more motivated on average than in other classes, and variation due to a rater effect because some teachers rate their students more positively or more negatively on average). To appropriately model these two sources of variation, both a shared (rater effect) construct and a configural construct (to mirror the individual-level construct) are needed, as shown in Figure 6. In this model, the factor loadings should be constrained across levels for all p variables ($\boldsymbol{\lambda}_W = \boldsymbol{\lambda}_B$) as part of the configural construct (ξ_{B1}). In addition,

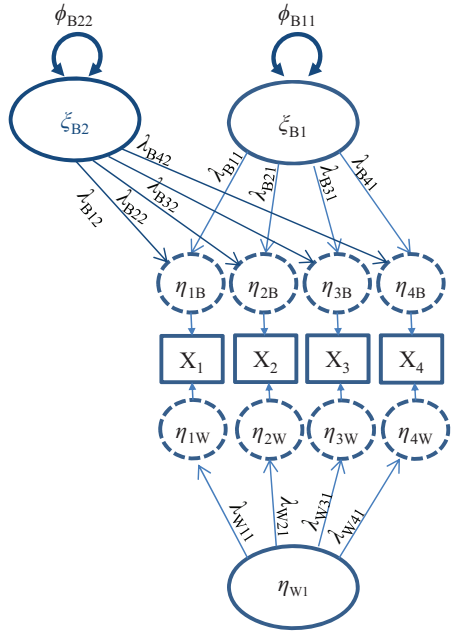

FIGURE 6. *Simultaneous shared and configural cluster construct model with required cross-level constraints* ($\lambda_{W11} = \lambda_{B11}$, $\lambda_{W21} = \lambda_{B21}$, $\lambda_{W31} = \lambda_{B31}$, $\lambda_{W41} = \lambda_{B41}$).
Note: Each item η *component also has residual with variance* θ_B *or* θ_W*, not shown for simplicity.*

one or more shared constructs (ξ_{B2}) can be modeled to explain additional covariation among cluster-average item responses.

Researchers should be aware, however, that what may appear to be an additional shared construct at the cluster level may be reflective of what we call a *spurious contextual effect*. Specifically, when constrained factor loadings across levels yield relatively poor fit as compared to a model with unconstrained loadings in a configural model, the loadings may differ across clusters due to measurement noninvariance across subpopulations within clusters, coupled with differences in the proportion of members in the subpopulations across the clusters. In the absence of auxiliary data to shed light on these possible subgroups, a strong theoretical rationale would be important in inferring if there is a true contextual effect (an additional shared construct) or a spurious contextual effect. As with the spurious ICC, a researcher cannot statistically evaluate which of these conditions (an additional shared construct or measurement noninvariance at the within-cluster level) are correct, unless, of course, it is known which subpopulation characteristic is associated with any noninvariance and indicators of that subpopulation are available. However, given the observational data, it can

never be known that the appropriate subpopulations have been identified. In Section 4 of this article, we provide a simple simulation to demonstrate this spurious contextual effect caused by measurement noninvariance across subpopulations within clusters.

2.3. Summary of Conceptual Issues

In this section, we have laid out five possible situations in the measurement of constructs using item-level responses from individuals in nested data settings: an individual-level construct with planned use (1) across a broad population or (2) within a cluster, (3) a shared cluster-level construct, (4) a cluster-level configural construct, and (5) a context with a simultaneous cluster-level shared and configural construct. The CFA models for structural validation for each of these are different, and careful thought is required to select the appropriate model, as opposed to what appears to be the conventional approach of applying MCFA simply because the data have a nested structure. Before considering the use of MCFA to provide structural validation evidence, a researcher should have a well-defined unit of interest and intended future use and hypothesized meaning of the construct(s) at the level(s) of interest. While strong theoretical rationale is a required condition to choose a proper CFA model, some model comparison approaches such as testing random variances or loadings, and possible multiple factors at the cluster level, can be utilized to capture various aspects of constructs as well as to validate the structural features of constructs.

3. Illustrative Examples

The following examples are intended to highlight some of the issues that an applied researcher might face when examining the structural validity of their proposed measure. For simplicity and consistency in display, in all examples, we present 4 possible items that might be used to measure a given construct; we subjectively selected these items from a larger pool of questionnaire items in the available data sets. We use two public-release data sets: Early Childhood Longitudinal Study of Kindergarten (ECLS-K; Tourangeau et al., 2009) and Education Longitudinal Study of 2002 (ELS:2002; Ingels, Pratt, Rogers, Siegel, & Stutts, 2004). Although both ECLS-K and ELS:2002 involved a fairly complex sampling structure, for simplicity, we assume here that the data were drawn from a simple random sample of schools and a simple random sample of students or teachers within each selected school. For advice on accommodating more complex sampling structures with multilevel modeling for inference to the U.S. population, see Asparouhov and Muthén (2006), Rabe-Hesketh and Skrondal (2006), and Stapleton and Kang (2016). We conduct six analyses, demonstrating assessment of the structural aspect of validity of a hypothesized individual construct (importance of social skills) both across a broad population and for use within clusters only, two hypothesized shared constructs (school violence and

TABLE 2.
Observed Distributions of the 4-Item Responses Across the Entire Sample

Item	Essential (%)	Very Important (%)	Somewhat Important (%)	Not Very Important (%)	Not Important (%)	ICC(1)
Shares	32.8	61.3	5.6	0.3	0.0	.04
Draws	22.1	51.0	24.0	2.5	0.4	.02
Is calm	24.3	58.3	16.1	1.1	0.2	.04
Communicates well	35.1	58.5	6.1	0.3	0.1	.04

Note. ICC = intraclass correlation.

positive school culture), and a configural dispersion model of the hypothesized importance of social skills construct. Finally, we show an example of a simultaneous shared and configural model of positive school culture (shared) and feelings of support (configural).

3.1. Examples for Individual as Unit of Interest

Using the ECLS-K data, we investigated the plausibility of a single factor underlying the responses to 4 items that tap a parent's belief about social skills needed for kindergarten. Specifically, parents of kindergartners were posed the following question stem during a phone interview in the fall: "Now I'm going to ask you how important you think it is for children to know or do certain things to be ready for kindergarten. How important do you think it is that a child . . . ?" Four items posed the following concepts: "shares," "draws," "is calm," and "communicates well." The available response options were *essential*, *very important*, *somewhat important*, *not very important*, and *not important*. We hypothesized that a single construct importance of social skills was underlying the responses to the 4 items. Such a construct would be potentially useful in identifying parents prior to kindergarten for intervention, such as targeted evening orientations by the school district, or information sent home to targeted parents of children in a school. Responses were obtained from 16,760 parents of students nested within 948 schools (for simplicity in this demonstration, missing data were treated by using listwise deletion for all analyses). Cluster sizes ranged from 1 to 27 students per school, with an average of 17.7 students per school. The response distribution on the 4 items is shown in Table 2 along with ICC values. One might make the assumption that the responses to these items were not subject to influence of the school environment (a reasonable assumption had the data been collected in early September; however, these data were collected between September and November). Assuming that the responses were not subject to influence of the school environment specifies that any

TABLE 3.
Model Fit Results From a One-Factor CFA of Individual Importance of Social Skills Construct

Fit Statistic/ Index	Single-Level Design-Based Adjusted	Single-Level Ignoring the Clustering	Within-Cluster Construct
Model χ^2	59.86 ($df = 2$)	87.12 ($df = 2$)	29.87 ($df = 2$)
CFI	0.99	0.99	1.00
RMSEA (90% CI)	.04 [.03, .05]	.05 [.04, .06]	.03
SRMR	.01	.01	.01 (within) .02 (between)

Note. CFA = confirmatory factor analysis; CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

dependency within school found on the parent beliefs item responses (as reflected in the positive, yet minor, ICC values) is a result of parents of similar attitudes living in the same area; therefore, if the parents were moved to another school by an exogenous force, they would be expected to retain their same attitudes regarding the importance of social skills to kindergarten. Given this hypothesis of the lack of school influence on these item responses, a single-level model would be appropriate and any dependency of the 16,760 responses within 948 clusters could be accommodated in the estimation of standard errors and model fit statistics using a design-based approach to estimation.

Using Mplus 7.1, a single-level model of the four responses as item indicators of a single construct was run using a linearization method to obtain appropriate standard errors and adjusted χ^2 statistics.[2] For comparison, a naïve analysis was also conducted, ignoring that the students were nested within schools. Model fit results from these two models are shown in the first two columns of Table 3 and the resulting parameter estimates are shown in Table 4. From the results in Table 3, one can see that the adjustment for the (assumed spurious) ICC via the linearization method resulted in fit statistics that suggest a better fitting model in terms of both the model χ^2 statistic and those indices that use the χ^2 statistic in their calculations.

Furthermore, it can be noted in Table 4 that some of the factor loading standard errors, as well as those for the item residual variances, are larger for the design-based estimation as compared to that which assumed independence of observations. From these results, one might conclude that a model with a single latent construct of social skills importance is reasonable, given the very strong indices of model fit: comparative fit index, root mean square error of approximation, and standardized root mean square residual (SRMR). Given the significance of the model χ^2 statistic, however, it can be assumed that the model is misspecified to some extent. Examining possible model modifications, the largest correlation between any of the item residuals would have been approximately 0.13 (*share* with *communicates well*), a value that is below typical criteria for

TABLE 4.
Parameter Estimates From a One-Factor CFA of Individual Importance of Social Skills Construct

	Single-Level Design-Based Adjusted		Single-Level Ignoring the Clustering		Within-Cluster Construct	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Unstandardized parameter estimates						
λ_{share}	1.00	—	1.00	—	1.00	—
λ_{draws}	1.17	0.026	1.17	0.023	1.18	0.027
λ_{calm}	1.05	0.025	1.05	0.022	1.05	0.025
λ_{commun}	0.90	0.020	0.90	0.017	0.89	0.021
χ_{social}	0.15	0.004	0.15	0.004	0.14	0.004
θ_{share}	0.18	0.004	0.18	0.003	0.18	0.004
θ_{draws}	0.39	0.007	0.39	0.006	0.38	0.007
θ_{calm}	0.30	0.006	0.30	0.004	0.29	0.006
θ_{commun}	0.23	0.005	0.23	0.003	0.22	0.005
Standardized loadings						
λ_{share}	0.67		0.67		0.66	
λ_{draws}	0.58		0.58		0.58	
λ_{calm}	0.59		0.59		0.59	
λ_{commun}	0.58		0.58		0.57	

Note. CFA = confirmatory factor analysis.

concluding local dependence (Yen, 1993). Using the item parameter estimates, then, we might then opt to calculate composite reliability ω, assuming a unidimensional scale. Using the first or second set of estimates in Table 4 in Equation 2, we obtain a reliability estimate of $\hat{\omega} = .70$ for a weighted composite based on the 4-item responses.

Note that, unlike above where the construct is intended to be used in the future across the population of students, the construct may instead be intended to be used in the future only within clusters. For example, a principal may want to differentiate parents of children in the school based on their higher or lower importance ratings on social skills. If this is the intention, then an alternate model (Model 2 in Table 1) would need to be evaluated, based on Figure 3. The model fit results for this within-cluster measurement model are shown in the third column of Table 3 and the parameter estimates are shown in the third column of Table 4. Similar to the single-level model fit, the fit appears good. Composite reliability is estimated just slightly lower at .69 because of attenuation, given that the between-cluster variance of the items has now been removed. Note that the unstandardized within-cluster loadings would be expected to be the same in columns 1 and 3 when no shared cluster-level construct is affecting item responses.

3.2. Examples for Clusters as Unit of Interest

In this section, we present two examples that examine cluster-level shared constructs and another two examples that investigate configural constructs (aggregates of individual-level constructs).

3.2.1. Shared construct examples. Our first example is intended to demonstrate an (unsuccessful) investigation of a cluster-level construct underlying the responses to 4 items asking students to relay their experiences in terms of school violence, using the ELS:2002 data. Specifically, 10th graders answered questions regarding how often they had experienced some violent incidents using the following 4 items: ''Had something stolen at school,'' ''Someone offered drugs at school,'' ''Someone threatened to hurt [10th grader] at school,'' ''Someone hit [10th grader].'' Suppose a researcher hypothesized that these 4 items tapped a school-level school violence construct that is believed to be shared among students within schools. Responses were obtained from 12,558 students nested within 748 schools. Cluster sizes ranged from 1 to 27 students, with an average of 16.7 students per school. The response distributions on the 4 items are shown in Table 5 along with the ICC values. ICC(1) values were obtained from the Mplus software output, while ICC(2) values were estimated using Equation 8.

Given the item ICC(1) values, it is difficult to justify the use of the current school violence measure as a shared construct among the students in the same school. This may indicate that such a construct does not exist or at least the information obtained from students using the current items is not appropriate to measure such a construct. In fact, examining the item wording should alert us to possible concerns in the measurement of a shared construct. For the items hypothesized to measure school violence, all item stems were self-referent, specifically asking students to reflect on their own experiences and one not necessarily shared by all students; item responses would not be expected to be isomorphic. We therefore decided that these items should not be considered useful for measuring a school-level shared construct.

Our second example investigates a cluster-level shared factor underlying the responses to 4 items that ask third-grade teachers' opinions about positive school culture using the ECLS-K data. Likert-type scale responses to the following 4 items were used to measure this shared construct: ''Staff have school spirit,'' ''Staff accept me as colleague,'' ''Staff learn/seek new ideas,'' and ''Parents support school staff.'' Responses were collected from 2,839 teachers nested within 919 schools. Cluster sizes ranged from 1 to 19 teachers, with an average of 3.1 teachers per school. The response distributions on the 4 items are shown in Table 6, along with the ICC values.

For structural validation purposes, the shared construct factor model (Model 3 in Table 1) shown in Figure 4 was posited for the positive school culture construct and

TABLE 5.
Observed Distributions of the 4-Item Responses for School Violence

Item	Never (%)	Once or Twice (%)	More than Twice (%)	ICC(1)	ICC(2)
Stolen	59.9	34.0	6.1	.05	.47
Offered drugs	77.2	14.0	8.8	.10	.65
Threatened	78.0	16.3	5.7	.06	.52
Hit	78.9	15.3	5.8	.05	.47

Note. ICC = intraclass correlation.

TABLE 6.
Observed Distributions of the 4-Item Responses for Positive School Culture

Item	Strongly Disagree (%)	Disagree (%)	Neither Agree nor Disagree (%)	Agree (%)	Strongly Agree (%)	ICC(1)	ICC(2)
Have school spirit	0.9	4.6	10.1	57.8	26.6	.29	.56
Accept as colleague	0.5	0.8	3.6	51.3	43.8	.12	.30
Learn new ideas	0.3	2.0	6.7	48.4	42.7	.20	.44
Parents support staff	0.3	5.5	15.0	58.8	20.4	.27	.53

Note. ICC = intraclass correlation.

TABLE 7.
Model Fit Results From Shared Construct Model of Positive School Culture

Fit Statistic/Index	Estimate
Model χ^2	1.517 ($df = 2$)
CFI	1.00
RMSEA	.000
SRMR	.02 (within) .018 (between)

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

the model fit results are reported in Table 7. The item ICC(1) values for manifest variables ranged from 0.12 to 0.29 and ICC(2) estimates ranged from 0.30 to 0.56. Depending on the number of raters who provide measures of the (hypothesized)

TABLE 8.
Between-Level Parameter Estimates From Shared Construct Model of Positive School Culture

Parameter	Unstandardized Estimate	*SE*	Standardized Estimate	*SE*
λ_{spirit_b}	1.000	—	0.417	0.027
λ_{accept_b}	0.477	0.057	0.199	0.022
λ_{learn_b}	0.650	0.058	0.271	0.024
$\lambda_{support\_b}$	0.595	0.071	0.248	0.026
$\chi_{positive\_b}$	0.174	0.023		
θ_{spirit_b}	0.010	0.012		
θ_{accept_b}	0.010	0.005		
θ_{learn_b}	0.027	0.007		
$\theta_{support\_b}$	0.096	0.013		

shared items, this example possibly satisfies the minimal standard for ICC(2) of 0.5 (Klein et al., 2000). For the item with the highest amount of clustering (*Staff have school spirit*), given cluster sizes ranging from 1 to 19 teachers per school, the reliability of any given school mean estimate on that item ranges from .29 to .89 (using Equation 8). It should be noted that 1 of the 4 items hypothesized to measure positive school culture was a self-referent item (*Staff accept me as a colleague*) and, in fact, had the lowest standardized factor loading and ICC values. Careful consideration of referent in items considered to measure shared constructs is necessary. From a measure development standpoint, if a shared construct is of interest, item referents should focus on the cluster and not the individual.

Because a measure of a shared construct is intended to be used at the cluster level, between-level composite reliability needs to be reported. The composite reliability estimate was calculated using Equation 2 with the Level 2 parameter estimates in Table 8 and was 0.90 for positive school culture. This large composite reliability is partially due to very small between-level residual variances. While values of ICC(1) and ICC(2) take within-level variance into account, the cluster-level composite reliability considers only between-level variance and between-level factor loadings. As a result, although ICC(1) and ICC(2) values, for the most part, are not terribly large in this example, the level-specific (between-cluster level) composite reliability estimate is high. Accordingly, researchers should address both quantities and report them to discuss validity and reliability of a shared cluster-level construct. Specifically, a researcher should be wary about using cluster-level constructs based solely on evidence of cluster-level composite reliability; it has been found to be positively biased under conditions with low ICCs (Geldhof et al., 2014).

3.2.2. Configural construct examples. We now present configural model examples using the items intended to tap importance of social skills and the items that

were used for positive school culture. Given our original example of the individual measure of importance of social skills, it could be that a researcher was interested in using a measure of this (individual level) construct as a configural construct at the cluster level; he wants to model with an aggregate construct regarding the parent opinions within schools. Knowing already that the ICC(1) values for the 4 items are very low (ranging from .02 to .04), we cannot expect substantial variability in the mean values of the items and thus the configural construct mean across the clusters, but a researcher may be interested in whether these items can capture differences in the dispersion of the individual-level factor across clusters. Perhaps future research might be undertaken examining whether schools in which there is a common parental understanding differ from schools with heterogeneity in parental beliefs. From a measurement validation perspective, we need to demonstrate that the items can measure that discrepancy in dispersion. Model 4 from Table 1, shown as Figure 5, was analyzed first with only the measurement structure imposed (with both unconstrained and constrained loadings across levels) and then with a random variance of the latent construct at Level 1.

Statistically, the unconstrained model had slightly better fit with change in scaled χ^2 of 11.63, $df = 3$, $p = .01$. Given the sample size, however, it was decided to retain the model with the constrained loadings for parsimony. The estimates from these models are contained in the first two pairs of columns of Table 9 and, as should be expected, are similar to the estimates obtained when the construct was modeled as within-cluster only (see Column 3 of Table 4). To add the random variance of importance of social skills to the model, Bayesian estimation was used with default priors (for variances, the default is an inverse gamma distribution (0, −1)), fixing (for identification) the average standard deviation of the latent construct within schools to 0.371 (or a value of 0.138 for variance, based on the estimate shown in either of the first two pairs of columns in Table 9) and allowing the model to estimate the variability of the standard deviation across clusters. This model resulted in loading and residual variances very similar to the model with fixed variance but did result in a finding of statistically significant ($p < .05$) variability across clusters with the variance in standard deviation of the latent construct within schools as 0.002. A 95% plausible range of the values suggest that 95% of the clusters would have standard deviations of importance of social skills between 0.28 and 0.46.

For another example of a configural construct, we revisit the positive school culture measure that we earlier hypothesized to be a shared construct but for which we found questionable support. Now, we hypothesize that the cluster-level construct is only the reflection of an individual construct (which we will call perception of support); the configural model shown in Figure 2 is needed to validate this hypothesized structure (Model 4 in Table 1). Two different models were fitted: a configural factor model in which cross-level

TABLE 9.
Unstandardized Parameter Estimates From Configural Models of Importance of Social Skills

	Constrained Loadings Model		Unconstrained Loadings Model		Constrained Loadings and Random Variance	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	Posterior *SD*
λ_{share_w}	1.000	—	1.000	—	1.000	—
λ_{draws_w}	1.172	0.027	1.201	0.028	1.167	0.019
λ_{calm_w}	1.052	0.025	1.066	0.026	1.048	0.014
λ_{commun_w}	0.898	0.020	0.892	0.021	0.890	0.013
χ_{social_w}	0.138	0.004	0.136	0.004	0.138[a]	—
θ_{share_w}	0.179	0.004	0.180	0.004	0.178	0.003
θ_{draws_w}	0.383	0.007	0.380	0.007	0.384	0.005
θ_{calm_w}	0.292	0.006	0.290	0.006	0.291	0.004
θ_{commun_w}	0.224	0.005	0.225	0.005	0.224	0.003
λ_{share_b}	1.000	—	1.000	—	1.00	—
λ_{draws_b}	1.172	0.027	0.541	0.182	1.167	0.019
λ_{calm_b}	1.052	0.025	0.588	0.194	1.048	0.014
λ_{commun_b}	0.898	0.020	1.024	0.130	0.890	0.013
χ_{social_b}	0.008	0.001	0.011	0.002	0.008	0.001
θ_{share_b}	0.006	0.001	0.003	0.002	0.006	0.001
θ_{draws_b}	0.004	0.002	0.006	0.002	0.003	0.001
θ_{calm_b}	0.009	0.002	0.012	0.002	0.009	0.002
θ_{commun_b}	0.006	0.001	0.002	0.002	0.005	0.001
$\mathrm{Var}(\sqrt{\phi_{social\_w}})$	—	—	—	—	0.002	0.0003

[a]Estimate is random across clusters.

measurement invariance is assumed (constrained loading model) and another model that allows cross-level measurement noninvariance (unconstrained loading model). The unconstrained factor loading model yielded better fit with a significant χ^2 difference test (change in scaled χ^2 was 41.13 with 3 *df*; see Table 10). Additionally, the between-level SRMR for the constrained model is substantial at 0.085. Although this phenomenon of configural model misfit can be seen from another perspective (e.g., possible measurement noninvariance across subpopulations at within clusters), we hypothesized that there is a shared aspect in the cluster-level variability that is reflected in the different factor loadings across levels. Accordingly, a hypothesized simultaneous shared and configural factor model (Model 5 from Table 1) as shown in Figure 6 was fit to the same response data and the results are summarized in Table 10 for robust model fit statistics and fit indices and Table 11 for parameter estimates.

TABLE 10.
Model Fit Results From Multilevel Configural Models of Positive School Culture (Perception of Support)

Fit Statistic/Index	Constrained Loadings Model	Unconstrained Loadings Model	Shared and Configural Factor Model
Model χ^2	60.760 ($df = 7$)	14.625 ($df = 4$)	13.869 ($df = 4$)
CFI	.96	.99	.99
RMSEA	.052	.031	.029
SRMR	.029 (within), .085(between)	.020 (within), .021(between)	.020 (within), .021(between)

Note. CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

TABLE 11.
Unstandardized Parameter Estimates From Multilevel Configural and Simultaneous Shared/Configural Model of Perception of Support/Positive School Culture

	Constrained Loadings Model		Unconstrained Loadings Model		Shared and Configural Construct Model	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
λ_{spirit_w}	1.000	—	1.000	—	1	—
λ_{accept_w}	0.834	0.061	1.106	0.102	1.088	0.099
λ_{learn_w}	1.034	0.067	1.306	0.112	1.281	0.113
$\lambda_{support\_w}$	0.686	0.044	0.725	0.066	0.723	0.062
$\chi_{positive\_w}$	0.174	0.014	0.125	0.015	0.128	0.016
θ_{spirit_w}	0.295	0.019	0.317	0.018	0.314	0.018
θ_{accept_w}	0.230	0.015	0.216	0.016	0.216	0.016
θ_{learn_w}	0.211	0.015	0.195	0.016	0.197	0.016
$\theta_{support\_w}$	0.366	0.016	0.373	0.016	0.373	0.016
$\lambda_{spirit\_b\_conf}$	1.000	—	1.000	—	1	—
$\lambda_{accept\_b\_conf}$	0.834	0.061	0.457	0.058	1.088	0.099
$\lambda_{learn\_b\_conf}$	1.034	0.067	0.628	0.059	1.281	0.113
$\lambda_{support\_b\_conf}$	0.686	0.044	0.613	0.069	0.723	0.062
$\chi_{positive\_b\_conf}$	0.090	0.018	0.179	0.023	0.014	0.012
$\lambda_{spirit\_b\_shared}$	—	—	—	—	1	—
$\lambda_{accept\_b\_shared}$	—	—	—	—	2.590	0.446
$\lambda_{learn\_b\_shared}$	—	—	—	—	1.434	0.188
$\lambda_{support\_b\_shared}$	—	—	—	—	1.514	0.269
$\chi_{positive\_b\_shared}$	—	—	—	—	0.026	0.010
θ_{spirit_b}	0.054	0.011	0.006	0.012	0	—
θ_{accept_b}	0.005	0.005	0.010	0.005	0.008	0.005
θ_{learn_b}	0.017	0.007	0.027	0.007	0.026	0.007
$\theta_{support\_b}$	0.104	0.013	0.094	0.013	0.095	0.013

This simultaneous shared and configural factor model fit is acceptable and the fit indices are very similar to the unconstrained loading model. As noted earlier, the variability of between-level residuals is very small and one of the items (''School has spirit'') yields a nonpositive residual variance if two factors are extracted at the cluster level. Therefore, the residual variance at Level 2 for this item was fixed at zero in this case. This model has thus provided support for the structural validity of a shared positive school culture construct, above and beyond the reflection of individual differences on the measure (the configural construct of perception of support). For the shared construct in this model, the composite reliability estimate is .90, using estimates found in Table 11.

In summary, in this section, we have presented six examples illustrating structural validation models with nested data. In each case, the researcher needs to identify the nature of the level of the construct prior to determining the appropriate model. Although only briefly touched upon here, each analysis involved several decisions and steps in terms of assessing fit and calculating appropriate indices of reliability, such as level-specific ω.

4. Simulation Demonstrations

In this section, we demonstrate two conditions where there appears to be a clustering effect, but it is actually reflecting a spurious relation caused by disproportional subpopulation membership across clusters coupled with subpopulation differences in the latent mean and/or measurement model. We intend for these small demonstrations to alert readers to the possible sources of variability that may be unrelated to the construct of interest but caused by the sample composition of clusters. For each demonstration, the following simple but illustrative situation is imposed:

1) There are four manifest items that reflect a unidimensional construct at the individual level.
2) The construct is individual only with no cluster construct that causes individual responses.
3) Two subpopulations exist within each cluster and, within each subpopulation, factor variance and total item variance are set to a value of 1.0.
4) Across the sample (ignoring cluster membership), students are distributed equally into the two subpopulations. This choice is made for convenience; results generalize to other distributions.
5) Data for 200 clusters are generated with each cluster containing 20 individuals.

Data were generated using SAS (Version 9.3) and ICC(1) values were calculated using analysis of variance components. All models were estimated in the Mplus (Version 7.1) software using maximum likelihood estimation with robust corrections to fit statistics and standard errors. To demonstrate how spurious

ICCs and spurious contextual effects can appear, we manipulated three parameters: the variability across clusters in the proportions in the subpopulations, the level of latent mean difference in the subpopulations, and the degree of loading noninvariance across subpopulations. First, we manipulated the proportion in Subpopulation 1 in the clusters (e.g., one half of clusters has 60% of the individuals in Subpopulation 1 and the other half of clusters has only 40% of the individuals in Subpopulation 1). Five conditions were examined: 50% versus 50%, 60% versus 40%, 70% versus 30%, 80% versus 20%, and 90% versus 10%. These distributions yield variances of the proportion of Subpopulation 1 (σ_π^2) of 0.0, 0.01, 0.04, 0.09, and 0.16, respectively. The variance of Subpopulation 1 of the 0.0 condition would be expected to result in ICC(1) values of 0.0, regardless of the level of latent mean differences across the subpopulations.

The second parameter manipulated was the level of latent mean difference ($\kappa_1 - \kappa_2$) between the two subpopulations. The differences in latent means were generated to be 0, 1, 2, or 3 units. Given that latent factors within a subpopulation were generated to have a variance of 1, the latent effect size (Hancock, 2001) of these differences can be seen as Cohen's $d = 0$, 1, 2, and 3 at the latent mean level and between 0 and 2.1 at the manifest item level depending on the item-factor loading. In this first simulation demonstration, loadings are equivalent across the two subpopulations at values of 0.5, 0.5, 0.7, and 0.7 for the 4 items, respectively. Expected mean differences across the two populations at the item level were thus 0, 0.5, 1, and 1.5 for Items 1 and 2 and 0, 0.7, 1.4, and 2.1 for Items 3 and 4. Using these two manipulated parameters, we demonstrate the issue of spurious ICC(s).

In Table 12, we show the ICC(1) values for each item, averaged across 1,000 replications. It can be shown algebraically that, in the balanced case, the expected value of the ICC(1) for a given item is a function of the variability of the proportion of individuals in Subpopulation 1 in clusters and the size of the difference in expected means across the two subpopulations:

$$\mathrm{E}[\mathrm{ICC}(1)] = \sigma_\pi^2(\mu_1 - \mu_2)^2 / \ (\sigma_\mu^2 + \sigma_w^2), \tag{9}$$

where μ_1 is the expected value of the item mean for Subpopulation 1, μ_2 is the expected value of the item mean for Subpopulation 2, σ_μ^2 is the variance of the subpopulation means and σ_w^2 is the subpopulation item variance, assumed constant across subpopulations. It is worth noting that as subpopulations become more disparate in their means, the total variability of the item (the denominator in Equation 9) increases, however, this variability does not manifest in between-cluster variability if the proportions of each subpopulation are equivalent across clusters ($\sigma_\pi^2 = 0$). For the numerator (between-cluster variability) to be positive, subpopulations must differ in their means and also the subpopulations must exist at differential rates across the clusters. As shown in the top half of Table 12, ICC(1) values that are typically used by researchers to assume that there is a clustering

TABLE 12.
Average ICC(1) Values by Variability in Proportion of Subpopulations Across Clusters and Subpopulation Means and Loadings Across 1,000 Replications

	Variability of Proportion of Subpopulation 1 Across Clusters				
Condition	$\sigma_\pi^2 = 0$	$\sigma_\pi^2 = .01$	$\sigma_\pi^2 = .04$	$\sigma_\pi^2 = .09$	$\sigma_\pi^2 = .16$
κ and λ same across subpopulation	.00	.00	.00	.00	.00
κ differs and λ same across subpopulation ($\kappa_1 - \kappa_2 = 1$, 2, or 3)					
Items 1 and 2					
$\mu_1 - \mu_2 = 0.5$	−.00	.00	.01	.02	.04
$\mu_1 - \mu_2 = 1.0$	−.01	.00	.02	.07	.12
$\mu_1 - \mu_2 = 1.5$	−.02	−.00	.04	.12	.22
Items 3 and 4					
$\mu_1 - \mu_2 = 0.7$	−.01	.00	.01	.04	.07
$\mu_1 - \mu_2 = 1.4$	−.02	.00	.04	.11	.20
$\mu_1 - \mu_2 = 2.1$	−.03	−.01	.06	.17	.33
κ same and λ differs across subpopulation ($\lambda_1 = 0.7$, 0.9 and $\lambda_2 = 0.3$, 0.5)	.00	.00	.00	.01	.02
κ and λ differ across subpopulation (higher κ has higher λ)					
Items 1 and 2					
$\mu_1 - \mu_2 = 1.1$	−.01	.00	.03	.08	.14
$\mu_1 - \mu_2 = 1.8$	−.02	.00	.04	.11	.20
$\mu_1 - \mu_2 = 2.5$	−.03	−.01	.07	.20	.38
Items 3 and 4					
$\mu_1 - \mu_2 = 1.3$	−.02	.00	.03	.10	.18
$\mu_1 - \mu_2 = 2.2$	−.02	.00	.05	.15	.28
$\mu_1 - \mu_2 = 3.1$	−.04	.00	.08	.23	.44

effect can easily be obtained if subpopulations exist and those subpopulations differ in their latent means and differ in their allocation across the clusters.

Our conclusion here is that, although the construct is a measure of an individual-level attribute only, just due to differential membership across clusters, there may appear to be a clustering effect. Using empirical values of the ICC(1) does not inform the analyst whether a construct is an individual one or one that is influenced by cluster membership.

In further examining sources of spurious ICC(1) values, we also chose to simulate the situation when there was measurement noninvariance across subpopulations. Data were generated based on just one set of parameters: The 4 items' respective loadings were 0.7, 0.7, 0.9, and 0.9 for one subpopulation and 0.3, 0.3, 0.5, and 0.5 for the other. We crossed this condition with the condition of latent mean differences of 0, 1, 2, and 3. Resultant ICC(1) values are presented in

the lower half of Table 12. As seen, the combination of latent mean differences with measurement noninvariance can result in even greater ICC(1) values, given the effects on estimated item means.

Of even more concern, however, is that model fit in a multilevel model can be more difficult to establish as compared to a single-level model. As part of our simulations, we ran three models: single level with a design-based correction to both the standard errors and the χ^2 model fit statistic, a two-level configural model with constrained cross-level loadings, and a two-level model with unconstrained loadings. For these same conditions shown in Table 12, where the ICC(1) is positive due to differential subpopulation membership within clusters and either subpopulation mean differences or measurement noninvariance, the model fit information is shown in Table 13 for the high variability in proportion distribution. The table for the other proportion variability condition that results in substantial ICC(1) values is in the supplement materials, available in the online version of the journal. In all cases, the single-level model with design-based correction performed well; the highest model rejection rate for this properly specified model was .07. It should be noted that parameter estimates, specifically loading estimates, from this single-level model reflect the marginal relationship based on a mixture of the two subpopulations.

Although the two-level models would be rejected more often than appropriate, this finding is well established in that the likelihood ratio tests based on maximum likelihood do not perform well when ICC(1) values are small or the number of individuals within clusters is below 50 (Hox & Maas, 2004; McNeish & Stapleton, 2014; Schweig, 2014). Importantly, it should be noted that if the multilevel model with constrained cross-level loadings fit as well as one with unconstrained loadings, then the only differences across the clusters were differences in means due to disproportionate Subpopulation 1 membership (see the first three lines of Table 13). The constrained model would have been (inappropriately) rejected in favor of the unconstrained model, at most at a rate of .07, close to a nominal α rate of .05, and thus generally would result in an appropriate conclusion. However, anytime there is measurement noninvariance across subpopulations, even when the latent means are the same, this noninvariance results in an increased likelihood of finding model misfit at the between-cluster level (at rates ranging from .15 to .61). The difference in the covariances of items across subpopulations due to measurement noninvariance is being absorbed into the between-cluster-level covariance matrix. In these cases, an analyst might incorrectly be tempted to claim that there is evidence of a shared construct in addition to the configural construct that reflects the individual attributes within cluster.

Although very limited in scope, these small simulations were intended to show the applied researcher that positive ICC(1) values and the rejection of MCFA models with cross-level loading constraints do not necessarily reflect the existence of a cluster-level construct. Simple explanations for those findings might possibly be found in examining the subpopulations that may exist differentially within clusters,

TABLE 13.
Model Fit Comparison Across 1,000 Replications, for the $\sigma_\pi^2 = .16$ Condition

Condition	Configural Constrained Model		Configural Un-Constrained Model		Proportion Unconstrained Model Preferred	Single-Level Design Adjusted Model	
	Average χ^2	Proportion Rejection	Average χ^2	Proportion Rejection	Via Scaled $\Delta\chi^2$ Test	Average χ^2	Proportion Rejection
κ differs and λ same across subpopulation							
$\kappa_1 - \kappa_2 = 1$	16.3	.41	35.8	.40	.07	2.0	.05
$\kappa_1 - \kappa_2 = 2$	15.8	.40	33.3	.40	.05	1.8	.04
$\kappa_1 - \kappa_2 = 3$	15.3	.42	40.0	.42	.07	2.2	.07
κ same and λ differs across subpopulation ($\lambda_1 = 0.7, 0.9$ vs. $\lambda_2 = 0.3, 0.5$)	47.2	.78	37.2	.39	.44	2.1	.06
Neither κ nor λ are the same across subpopulation (higher κ has higher λ)							
$\kappa_1 - \kappa_2 = 1$	37.6	.85	40.1	.41	.61	2.2	.06
$\kappa_1 - \kappa_2 = 2$	20.9	.52	42.6	.46	.29	2.0	.05
$\kappa_1 - \kappa_2 = 3$	23.3	.64	31.6	.43	.15	2.1	.06

however, one can never know whether the appropriate subpopulations were identified. Applied researchers should be very careful to consider theory when hypothesizing that a cluster-level construct underlies item responses and seek auxiliary data regarding subpopulation membership that may help to inform whether differential membership across clusters may be leading to the spurious clustering effect.

Discussion

In this article, we presented the decisions faced in the structural validation of measures in a nested data context and provided options for validation of the structure of a measure. First and foremost, the analyst must determine whether interest lies in measurement at the individual level, cluster level, or possibly both. As part of this decision regarding the level of measurement, an argument for that level of measurement must be provided. If the measure is presumed to be at the individual level only, the analyst should provide supporting data or make a conceptual argument that any dependency in the measure reflects a spurious ICC and is not a result of the influence of a true, shared, cluster construct. Although not investigated here, when subpopulation data are not available, it may be possible to use finite mixture models (Lubke, 2010) to investigate subpopulations that may be leading to a spurious clustering effect. Alternately, if subpopulation data are available, researchers might examine the ICC(1) values for measures based on conditional item residuals (conditional on subpopulation membership). If these residual ICC(1) values are near 0, then a single-level multiple indicator multiple cause (MIMIC) model would address cluster dependency appropriately. In fact, we ran a single-level MIMIC model on our simulated data for the condition with measurement invariance across sub-populations but with latent mean differences of $\kappa = 3$ and the highest proportion of variability of subpopulation membership across cluster. Across 1,000 replications, the model χ^2 was on average 5.02 with degrees of freedom of 5; this result is similar to that obtained with a design-based adjustment (average χ^2 of 5.03) suggesting that no dependency remained to be addressed. It is impossible, however, for the researcher to know whether there are informative unobserved subpopulations that have not been accounted for in the modeling.

Conversely, if the measure is of interest at the cluster level, an argument should be made for any cluster-level measure regarding whether it is just a reflection of individual differences of persons within clusters (configural) or whether it reflects a single characteristic of the cluster (shared). If a shared construct, then an argument should be made regarding how the item wording is targeting the cluster characteristic and how individuals are believed to be interchangeable in its measurement (Bliese, 2000). If a researcher argues that a construct is a configural construct, then the analyst must clearly lay out the measures of interest regarding the construct, whether they be central tendency or dispersion and provide an explanation of why this configuration is important

to measure and how the measure might be used in the future. When a mix of both shared and configural constructs might exist at the cluster level, expectations of the degree of composition should be expressed as well as why, given item wording, there are some elements of both shared and configural constructs. During item development and testing phases, those items that are not clearly reflecting desired shared or configural constructs should be addressed.

For reporting, the analyst has several options regarding the statistics to be provided regarding the items and hypothesized construct; guidelines regarding what should be reported for MCFA are in development (Kim, Dedrick, Cao, & Ferron, under review), but we provide some tentative direction here. Descriptive statistics that indicate the amount of clustering for the manifest variables, ICC(1) and ICC(2), are helpful to describe the context, along with information about the number of individuals per cluster, as this would impact the ICC(2) estimate for any given cluster. For any proposed models, the researcher should evaluate fit of the model as well as parameter estimates in order to examine whether resulting estimates correspond to what has been theorized. Once the model is established, estimates of composite reliability (ω) can be calculated, but the trustworthiness of this measure depends on the data conditions in terms of size of clusters and level of ICC(1) (Geldhof et al., 2014). It was observed in our empirical demonstrations that cluster-level composite reliability estimates can be large although ICC(1) and ICC(2) values are not sufficiently large to provide support for a cluster-level shared measure. Caution is required to interpret cluster-level measures under these conditions. Additional guidance on the approaches for validation at each level is given by van Mierlo, Vermunt, and Rutte (2009). Further research, building on Geldhof et al. (2014), would be helpful to clarify guidelines in terms of the levels of ω needed to support valid inference at the cluster level, especially in light of lower values of ICCs.

One additional interesting aspect of cluster-level constructs is that, given the same items, the meaning of the construct is not necessarily stationary in its status as a configural or shared construct. For example, a simple reflection of an individual construct such as achievement can be seen as a configural construct at the school level, especially if measured at the start of a school year, perhaps on the first day of kindergarten. However, if the individual becomes fully immersed in a cluster culture, such as when the students in the school come to know that the school has high (or low) expectations of achievement, between-cluster variability would increase and any covariation among items would reflect both a configural construct and a shared construct. Documentation of the ability of a measure to move from configural only to both configural and shared can be used to demonstrate a measure's sensitivity in measuring cultural shifts (or malleability of the construct over time). Further research is needed to investigate such model performance for validation.

There are possible extensions or alterations to these models that could be considered for data that are not self-report or obtained from other types of research design. For example, with teacher ratings of students' characteristics

within classrooms, given the same rater for each classroom, there would be an expected shared "rater" effect at the cluster level, along with any individual-level (and configural) construct. Our investigation here has been on cross-sectional data, but repeated measures would also provide interesting opportunities for the structural validation of trait-state CFA models (Kenny & Zautra, 2001) that are not currently examined in a multilevel framework. Additionally, it should be noted that our models assumed that items were fixed, as typical with questionnaire data; extending the framework beyond this scenario would involve additional considerations.

The models presented here represent relatively simple extensions of current thinking in MCFA, and our examples have been correspondingly simple. Likewise, the structural validity discussed here is but one aspect of validity of a measure. More research is needed on the conditions under which each of these models yield robust estimates and appropriate inference to increase their likelihood of being used for instrument development for both individual- and cluster-level constructs as well as the required validation evidence that should accompany such hypothesized constructs.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Evaluating model fit in a multilevel structural model is not a simple exercise and research in this area is ongoing (see Ryu & West, 2009; Schweig, 2014).
2. This model assumed that the responses were continuous and multivariate normal, which is clearly violated, given the categorical response options. Thus, the model was also analyzed treating the items as ordered categorical with an adjusted weighted least squares estimator, and no differences in inference regarding measure structure were found.

References

Asparouhov, T., & Muthén, B. (2005). Multivariate statistical modeling with survey data. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*. Retrieved September 26, 2006, from http://www.fcsm.gov/05papers/Asparouhov_Muthen_IIA.pdf

Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. In *Proceedings of the American Statistical Association* (pp. 2718–2726). Seattle, WA: American Statistical Association.

Asparouhov, T., & Muthén, B. (2010). *Resampling methods in Mplus for complex survey data*. Retrieved October 2, 2011, from http://www.statmodel.com/download/resampling_methods5.pdf

Asparouhov, T., & Muthén, B. (2012). *General random effect latent variable modeling: Random subjects, items, contexts, and parameters*. Retrieved February 27, 2014, from http://www.statmodel.com/download/NCME_revision2.pdf

Bliese, P. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Brennan, R. L. (1995). The conventional wisdom about group mean scores. *Journal of Educational Measurement*, *32*, 385–396.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Press.

Casellas, F., Varela, E., Aparici, A., Casaus, M., & Rodriquez, P. (2009). Development, validation, and applicability of a symptoms questionnaire for lactose malabsorption screening. *Digestive Diseases and Sciences*, *54*, 1059–1065. doi:10.1007/s10620-008-0443-3

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

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

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design and analysis*. Occasional paper of the Stanford Evaluation Consortium, Stanford University, Stanford, CA.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: John Wiley.

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

Forer, B., & Zumbo, B. D. (2011). Validation of multilevel constructs: Validation methods and empirical findings for the EDI. *Social Indicators Research*, *103*, 231–265.

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

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373–388.

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

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

Hox, J., & Maas, C. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In K. G. Joreskog & D. Sorbom (Eds.), *Recent developments in structural equation models* (pp. 135–149). AK Houten, the Netherlands: Springer.

Ingels, S. J., Pratt, D. J., Rogers, J. E., Siegel, P. H., & Stutts, E. S. (2004). *Education Longitudinal Study of 2002: Base Year Data File User's Manual. NCES 2004-405*. Washington, DC: National Center for Education Statistics.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research*, *47*, 267–292.

Kenny, D. A., & Zautra, A. (2001). The trait-state models for longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 243–263). Washington, DC: American Psychological Association.

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (under review). Multilevel factor analysis: A review of reporting practices. *Multivariate Behavioral Research*.

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

Klein, K. J., Bliese, P. D., Kozlowshi, S. W. J., Dansereau, F., Gavin, M. B., Griffin, M. A., . . . Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 512–553). San Francisco, CA: Jossey-Bass.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multi-level 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.

Levy, R., & Choi, J. (2013). Bayesian structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed., pp. 563–623). Charlotte, NC: Information Age Publishing.

Li, F., Duncan, T. E., Harmer, P., Acock, A., & Stoolmiller, M. (1998). Analyzing measurement models of latent variables through multilevel confirmatory factor analysis and hierarchical linear modeling approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*, 294–306.

Lubke, G. H. (2010). Latent variable mixture models. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 210–219). New York, NY: Routledge.

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*, 444–467.

Marsh, H. W., Ludtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Koller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*, 106–124.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis and alpha factor analysis. *British Journal of Mathematical Psychology*, *23*, 1–21.

McDonald, R. P. (1978). Generalizability in factorable domains: Domain validity and generalizability. *Educational and Psychological Measurement*, *38*, 75–79.

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

McNeish, D., & Stapleton, L. M. (2014). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*. Advance online publication. doi:10.1007/s10648-014-9287-x

Meade, A. W., & Eby, L. T. (2007). Using indices of group agreement in multilevel construct validation. *Organizational Research Methods*, *10*, 75–96.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*, 259–284.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education/Collier Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into scoring meaning. *American Psychologist*, *9*, 741–749.

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

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods & Research*, *18*, 473–504.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series B*, *60*, 23–56.

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

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

Raykov, T. (2009). Estimation of maximal reliability for multiple-component instruments in multilevel designs. *British Journal of Mathematical and Statistical Psychology*, *62*, 129–142.

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*, 130–141.

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

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

Schweig, J. (2014). Multilevel factor analysis by model segregation: New applications for robust test statistics. *Journal of Educational and Behavioral Statistics*, *39*, 394–422.

Scrimshaw, N. S., & Murray, E. B. (1988). The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. *American Journal of Clinical Nutrition*, *48*, 1142–1159.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*, 28–58.

Stapleton, L. M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 183–210.

Stapleton, L. M. (2013). Using multilevel structural equation modeling techniques with complex sample data. In G. R. Hancock & R. O. Mueller (Eds), *Structural equation modeling: A second course* (2nd ed., pp. 521–562). Charlotte, NC: Information Age Publishing.

Stapleton, L. M., & Kang, Y.-J. (2016). Design effects of multilevel estimates from national probability samples. *Sociological Methods and Research*. Advance online publication. doi:10.1177/0049124116630563

Tourangeau, K., Nord, C., Le, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). *Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User's Manual for the ECLS-K Eight-Grade and K-8 Full Sample Data Files and Electronic Codebooks*. Washington, DC: National Center for Education Statistics.

van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*, 368–392.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213.

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

Authors

LAURA M. STAPLETON is an associate professor in the measurement, statistics, and evaluation program in the Department of Human Development and Quantitative Methodology at University of Maryland, 1229 Benjamin Building, University of Maryland, College Park, MD 20742; e-mail: lstaplet@umd.edu. Her research interests include analysis of survey data obtained under complex sampling designs and multilevel latent variable models, including tests of mediation within a multilevel framework.

JI SEUNG YANG is an assistant professor in the measurement, statistics, and evaluation program in the Department of Human Development and Quantitative Methodology at University of Maryland, 1225 Benjamin Building, University of Maryland, College Park, MD 20742; e-mail: jsyang@umd.edu. Her research interests include multilevel latent variable models and estimation.

GREGORY R. HANCOCK is a professor and director of the program in measurement, statistics, and evaluation in the Department of Human Development and Quantitative Methodology at the University of Maryland, 1230 Benjamin Building, College Park, MD 20742; e-mail: ghancock@umd.edu. His research interests include structural equation models, latent growth models, latent variable experimental design, power in latent variable models, and model-based reasoning.

Manuscript received June 19, 2015
Revision received December 7, 2015; February 21, 2016
Accepted March 28, 2016