# Modeling cluster-level constructs measured by individual responses

*Configuring a shared approach*

Jak, S.; Jorgensen, T.D.; ten Hove, D.; Nevicka, B.

**DOI**
10.1177/25152459231182319

**Publication date**
2023

**Document Version**
Final published version

**Published in**
Advances in Methods and Practices in Psychological Science

*General Article*

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
July-September 2023, Vol. 6, No. 3, pp. 1–18

DOI: 10.1177/25152459231182319
www.psychologicalscience.org/AMPPS

Sage

# Modeling Cluster-Level Constructs Measured by Individual Responses: Configuring a Shared Approach

**Suzanne Jak[1], Terrence D. Jorgensen[1], Debby ten Hove[2], and Barbara Nevicka[3]**

[1]Methods and Statistics, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands; [2]Department of Educational and Family Studies, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; and [3]Department of Work and Organizational Psychology, University of Amsterdam, Amsterdam, The Netherlands

## Abstract

When multiple items are used to measure cluster-level constructs with individual-level responses, multilevel confirmatory factor models are useful. How to model constructs across levels is still an active area of research in which competing methods are available to capture what can be interpreted as a valid representation of cluster-level phenomena. Moreover, the terminology used for the cluster-level constructs in such models varies across researchers. We therefore provide an overview of used terminology and modeling approaches for cluster-level constructs measured through individual responses. We classify the constructs based on whether (a) the target of measurement is at the cluster level or at the individual level and (b) the construct requires a measurement model. Next, we discuss various two-level factor models that have been proposed for multilevel constructs that require a measurement model, and we show that the so-called doubly latent model with cross-level invariance of factor loadings is appropriate for all types of constructs that require a measurement model. We provide two illustrations using empirical data from students and organizational teams on stimulating teaching and on conflict in organizational teams, respectively.

## Keywords

cluster-level constructs, multilevel confirmatory factor analysis, doubly latent model, configural constructs, shared constructs

Received 12/1/22; Revision accepted 5/27/23

Researchers frequently use the responses of individuals in clusters to measure constructs at the cluster level. For example, student evaluations may be used to measure the teaching quality of instructors, patient reports may be used to evaluate social skills of therapists, and residents' ratings may be used to evaluate neighborhood safety. In these three examples, the target construct is something that (in theory) varies only at the cluster level; the individuals within one cluster all share the same instructor, therapist, or neighborhood.

A contrasting type of cluster-level constructs can be defined as constructs that theoretically differ across individuals within the same cluster. Examples are reading skills of students in a classroom, depressive symptoms of individual patients of a therapist, and number of years that individuals live in a particular neighborhood. Although the target construct here is defined at the individual level, it is quite likely that the averages across clusters will also vary because of cluster-level factors. For instance, the average reading skills of students in different classrooms may differ because of differences of teaching styles across classrooms, the average amount of depressive symptoms of patients may differ across therapists because of differences in therapists' approaches,

**Corresponding Author:**
Suzanne Jak, Methods and Statistics, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands
Email: S.Jak@uva.nl

and some neighborhoods may have a higher turnaround of residence than other neighborhoods because of differences in local council policies and amenities.

When multiple items are used to measure cluster-level constructs with individual-level responses, multilevel confirmatory factor models are useful (Muthén, 1989, 1994). These models allow for the evaluation of the factor structure at the cluster level (modeling the variances or covariances among item means across clusters) and at the individual level (modeling the variances or covariances across individuals within clusters). How to model constructs across levels is still an active area of research (Bardach et al., 2020; Jak & Jorgensen, 2017; Morin et al., 2022; Stapleton et al., 2016; Stapleton & Johnson, 2019) in which competing methods are available to capture what can be interpreted as a valid representation of cluster-level phenomena. Moreover, the terminology used for the cluster-level constructs in such models varies across researchers (see Table 1 in the following section).

In this article, we therefore provide a comprehensive framework for classifying and modeling cluster-level constructs that are operationalized through individual-level responses. We start with an overview of different terminology used for multilevel constructs, and we classify these constructs based on whether (a) the target of measurement is at the cluster level or the individual level and (b) the construct requires a measurement model. Next, we discuss various two-level factor models that have been proposed for multilevel constructs that require a measurement model, and we show that the so-called doubly latent model with cross-level invariance of factor loadings is appropriate for all types of constructs that require a measurement model. We provide two illustrations using empirical data from students and from organizational teams and end with a reflection on some issues related to modeling cluster-level constructs with individual-level responses.

## Terminology for Cluster-Level Constructs

As indicated in the introduction, one can differentiate between (a) cluster-level constructs that target a cluster-level attribute that is shared among the individuals within a cluster and (b) constructs that target an individual attribute, which can be decomposed into individual- and cluster-level components. Moreover, one can think of constructs that are easily observable or directly measurable, in contrast with latent constructs that are only indirectly observable so that they need a measurement model. Crossing these two factors leads to four types of constructs. For ease of discussion, we use the example setting of students who are nested within classrooms and who share one teacher per classroom. In this example, one can imagine cluster-level constructs that are easy to observe or directly measure, such as a teacher's years of teaching experience or the percentage of boys in the classroom, and constructs that are harder to operationalize, such as teaching skills of the teacher and average student achievement in the class. The fourth column in Table 1 provides an overview of the labels that researchers have used to refer to these four example constructs. We discuss these types one by one.

### *Targeting the cluster and no measurement model is required: global constructs*

Years of teaching experience (see first row of Table 1) is a cluster-level construct that targets the cluster and is objectively quantifiable. One can directly ask the teacher, who is in the best position to provide an accurate measure of this construct. For these types of objective constructs, it would not be sensible to ask all the students to report on this variable, nor would it be necessary to use multiple items to operationalize the construct (i.e., no measurement model is required). Constructs such as years of experience, gender, or age of the teacher would be considered "global constructs" by Klein and Kozlowski (2000). They defined global constructs as follows (using the term "unit" where we use "cluster"):

> Global constructs pertain to the relatively objective, descriptive, easily observable characteristics of a unit that originate at the unit level. Global unit properties do not originate in individuals' perceptions, experiences, attitudes, demographics, behaviors or interactions, but are a property of the unit as a whole. (p. 29)

These authors also stressed that within-clusters variability should not exist for global constructs:

> There is no possibility of within-units variation because lower-level properties are irrelevant; indeed, any within-units variation is most likely the result of a procedure that uses lower-level units to measure the global property. If, for example, group members disagree about the size of their group, someone has simply miscounted. Unit size has an objective standing apart from members' characteristics or social-psychological processes. In contrast, "perceived group membership" is an entirely different type of construct. (Klein & Kozlowski, 2000, p. 30)

In global constructs, there should thus not be any variation at the individual level. If responses are gathered at the cluster level (which is to be expected), there is no variation possible at the individual level. If responses are gathered through individuals, the variation

**Table 1.** Overview of Four Theoretically Different Types of Cluster-Level Constructs and Their Characteristics

| Target of measurement | Measurement model needed? | Example | Terminology | Theoretical variation possible at | Source of cluster-level variation | Source of individual-level variation |
|---|---|---|---|---|---|---|
| **Cluster** | No | Years of experience of teacher | **Global construct**[a]<br>Truly-shared construct[b,c]<br>True level-two measure[d] | Cluster level only | Construct under study | None |
| | Yes | Student ratings of teacher quality | **Shared construct**[a]<br>Not-truly shared construct[b,c]<br>Climate construct[d]<br>Reflective construct[e] | Cluster and individual levels | Shared perceptions of construct under study | Individual differences in experiences, attitudes, perceptions, values, cognitions about cluster-level construct |
| **Individual** | No | % of boys in a classroom | Configural construct[a] | Cluster and individual levels | Aggregate of individual-level construct | Individual differences in construct |
| | Yes | Class average of students' achievement | **Configural construct**[a,c]<br>Contextual construct[d]<br>Formative construct[e] | Cluster and individual levels | Aggregate of individual-level construct | Individual differences in construct |

Note: Terms in bold are the terminology used in this article for this type of construct.
[a]Klein and Kozlowski (2000). [b]Bliese (2000). [c]Stapleton et al. (2016). [d]Marsh et al. (2012). [e]Lüdtke et al. (2008).

at the individual level must be measurement error. Because one does not need a measurement model for global constructs, we do not discuss the modeling of global constructs in this article.

## *Targeting the cluster and a measurement model is required: shared constructs*

Teaching quality is an example of a construct that targets the cluster level but is not directly quantifiable (see second row in Table 1). Teaching quality could, for example, be measured through students' ratings of their teachers using multiple items. The literature uses the term "shared" when a cluster-level characteristic is operationalized by asking individuals to report on the cluster-level construct. The definition of shared constructs is a bit ambiguous because some authors use this terminology for constructs that could also be considered global constructs. That is, some authors use the term "shared" for constructs that theoretically exclude the possibility of within-clusters variation, whereas other authors refer to shared constructs as constructs for which within-clusters variation can be expected. Because these differences in the definition of shared constructs have consequences for the type of statistical models that would be appropriate, we discuss the different positions on within-clusters variation below.

### *Position 1: within-clusters variation should not exist for shared constructs*

Bliese (2000) considered a construct to be shared only if there is complete within-clusters agreement. For a shared construct, the individual responses to the items should be interchangeable with perfect (interrater) reliability (IRR). In other words, for such a measure to be valid, individuals within a cluster should respond in an identical way. In the example of teaching quality, according to this position, all students rating the same teacher would completely agree about the teaching quality. Stapleton et al. (2016) adhered to this definition and stated that "there should be minimal variability found at the within-cluster level for a truly shared construct" (p. 492). These authors mentioned that criteria (e.g., high IRR) can be used to decide whether a construct can be considered shared. So their definition of shared constructs pertains to objective cluster properties about which all individuals in a cluster should theoretically give the same responses (when the measurements used are valid). If individuals do not agree, the construct is

not "truly" shared (at least, measurements are not composed only of the truly shared factor), and Bliese and Stapleton et al. argued that it is questionable whether the construct can be considered a shared one. The strict definition of shared constructs by Bliese and Stapleton et al.—which involves no variability in the responses at the within-clusters level—actually corresponds to the definition of global constructs by Klein and Kozlowski (2000). Thus, Bliese and Stapleton et al. did not make a distinction between global and shared constructs, whereas Klein and Kozlowski did.

### *Position 2: within-clusters variation can exist for shared constructs*

Klein and Kozlowski (2000) defined shared constructs as constructs that are shared by the members of a cluster and for which there should be some level of within-clusters agreement among cluster members. Shared constructs originate from the individual cluster members' experiences, attitudes, perceptions, values, cognitions, or behaviors and converge among cluster members. Klein and Kozlowski mentioned organizational climate, collective efficacy, and group norms as examples. According to this definition, shared constructs are thus constructs that target the cluster, for which ratings of the individuals in the cluster may differ. Given evidence of adequate agreement within clusters (e.g., high IRR), the aggregate value of the measure can be assigned to the cluster. This way, effectively, the shared perceptions of individuals in a cluster are interpreted as a proxy for the cluster-level attribute.

In the context of educational research, Marsh et al. (2012) introduced the term "climate constructs" for shared constructs. These authors acknowledged that although in a school context all students within the same class may be rating the same classroom climate, there may still be systematic differences among the ratings by students within each class:

> From this perspective, classroom climate is based on the shared perceptions among different students within the same class, whereas differences among students within the same class (residual L1 [i.e., individual-level] differences after controlling for shared agreement) are a source of unreliability in the L2 [i.e., cluster-level] climate construct. This is not to say that there are no systematic individual differences among the ratings by L1 students within each class, but merely that these individual differences do not reflect the L2 classroom climate of interest (i.e., the shared agreement among students from the same class). This point was made in the classic 1976 article by Cronbach, who noted that studying individual differences in the perceptions of different students within the same classroom might be interesting but does not reflect classroom climate. From this conceptual perspective, it follows that if there is no agreement among students within the same class in relation to the classroom climate variable, then the aggregated measure of climate is completely unreliable and probably should not be considered further. (p. 110)

From this perspective, the more agreement there is within the clusters, the more reliably the shared construct is measured, but the agreement does not have to be perfect (as would be the case for global constructs).

### *Definitions and terminology used in this article*

We adhere to the position stating that shared constructs are likely to show variation at the individual level. The individual-level variation represents the differences in individual cluster members' experiences, attitudes, perceptions, values, cognitions, or behaviors. In the remainder of this article, we adhere to the definitions of global and shared constructs as proposed by Klein and Kozlowski (2000).

### *Targeting the individual: configural constructs*

The proportion of boys in a classroom is an example of a cluster-level construct that targets the individual and for which one would not need a measurement model (see third row in Table 1). Although one needs individual information to be able to aggregate the individual scores to the cluster level, it is not necessary to obtain that information from the individuals themselves. For example, the school administration can provide information about the proportion of boys in each classroom. Individual-level variation in such configural constructs likely exists, reflecting within-clusters differences in the construct of interest. Because one does not need a measurement model for directly quantifiable configural constructs, we do not further consider these types of constructs in this article, reserving the term "configural construct" to the type explained in the next section (i.e., not directly quantifiable constructs that target the individual).

The average mathematical achievement of students in a classroom or the average job satisfaction in organizational teams are examples of cluster-level constructs that target the individual and that are not directly quantifiable (see last row of Table 1). Researchers will operationalize such constructs using multiple indicators. Configural constructs that are measured with multiple indicators, so that they need a measurement model, are also referred to as "contextual constructs" (Marsh et al., 2012) or

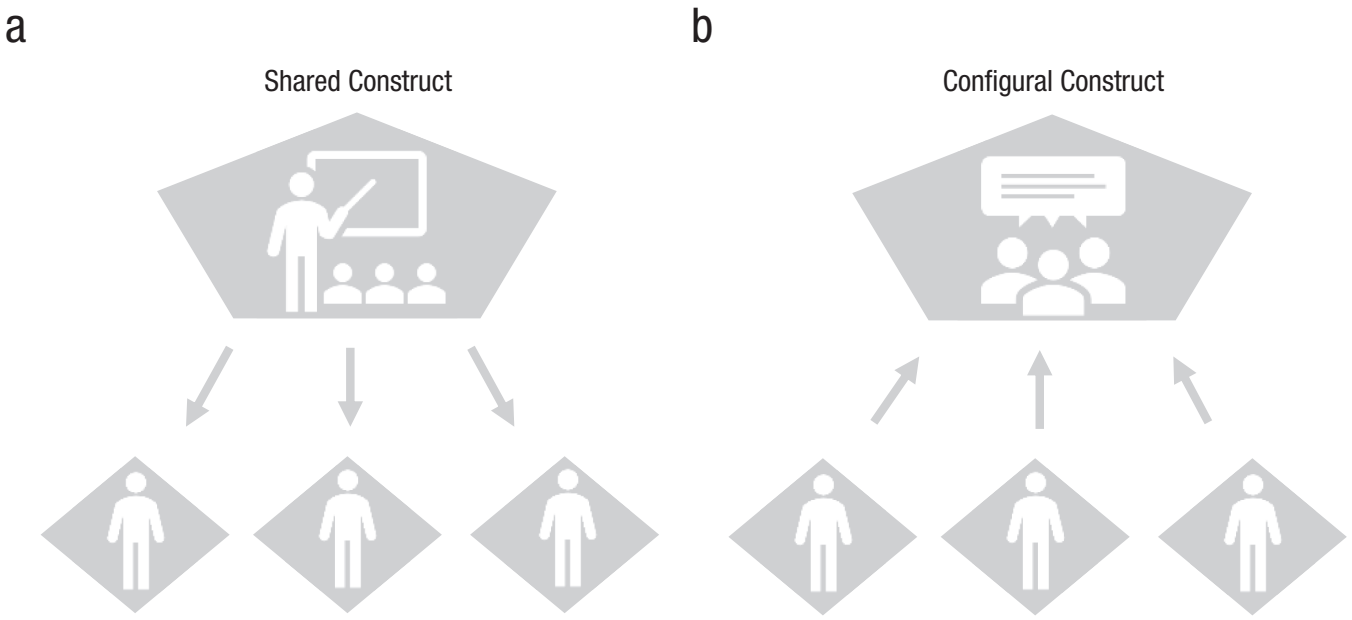


**Fig. 1.** Conceptual depiction of the relationship between the cluster-level construct and the individuals in the cluster for (a) shared constructs and (b) configural constructs. Given the similarity to the relation of items and latent variables in a reflective measurement model and formative model (see Edwards & Bagozzi, 2000), shared constructs are also denoted as "reflective constructs," and configural constructs are also denoted as "formative constructs" (Lüdtke et al., 2008).

"formative constructs" (Lüdtke et al., 2008). Stapleton et al. (2016) defined configural constructs as constructs based on an aggregate of the measurements of individuals who comprise the cluster. This matches the configural construct as defined by Klein and Kozlowski (2000), who stated that configural properties capture the array, pattern, or configuration of individuals' characteristics within a unit. The configural constructs are often based on the cluster averages, but configural constructs can also be represented by other quantities, such as the dispersion (standard deviation), minimum, maximum, or median values of the individuals in the cluster.

In this respect, the contextual constructs in the examples provided by Marsh et al. (2012) represent a specific type of configural construct, that is, a configural construct that is formed using the cluster averages of individual responses in which the target of the measurement is the individual. Stapleton et al. (2016) described modeling cluster differences in dispersion and in means, but Stapleton and Johnson (2019) later focused only on cluster averages of individuals, similar to Marsh et al. (2012). In this article, we too focus only on using cluster averages to represent configural cluster-level constructs. Configural constructs are likely (and intended) to show variation at the individual level. The individual-level variation represents the individual cluster members' differences on the construct of interest, and the cluster-level component is merely an aggregate that captures average between-clusters differences.

### *Other terms for configural and shared constructs*

In the terminology of Marsh et al. (2012), the difference between climate or contextual constructs is dictated by the item referent. If the referent of the item is the individual (e.g., "I like going to school"), then the cluster aggregate represents a contextual (i.e., configural) construct. If the referent of the item is the cluster (e.g., "My school is fun to go to"), then the cluster aggregate represents a climate (i.e., shared) construct.

Lüdtke et al. (2008) referred to configural constructs as formative constructs. The terms "formative" and "reflective" used by Lüdtke et al. are based on the concept of formative and reflective indicators of constructs (see Bollen & Diamantopoulos, 2017; Diamantopoulos & Siguaw, 2006; Edwards & Bagozzi, 2000). With reflective indicators, one single latent variable is assumed to cause the item responses, similar to how a shared cluster construct (e.g., how fun a school is to attend) causes the responses of the individuals in the cluster. With formative indicators, the item responses together form or cause the construct, similar to how responses of individuals (e.g., how much they like going to school) in a cluster can comprise a formative cluster-level construct (e.g., the average amount of enjoyment students in a school experience). Figure 1 shows a conceptual graphical display of the two types of aggregation processes described by Lüdtke et al. In the remainder of the article,

we use the term "configural constructs" to refer to formative or contextual constructs. The term "configural" is also used in the measurement-invariance literature, in which a "configural model" refers to a model in which the pattern of free and fixed factor loadings is equal across groups (Horn et al., 1983). The meaning of "configural" in that context has nothing to do with the configural constructs as defined by Klein and Kozlowski (2000). To avoid confusion, we refer only to configural *constructs* and not to configural *models*.

## *Constructs that are not easily classified as configural or shared*

So far, we have distinguished two types of cluster-level constructs that require a measurement model: shared constructs that target a cluster-level attribute and configural constructs that target an individual-level attribute. However, it may not always be possible to directly classify constructs as either configural or shared. Klein and Kozlowski (2000) illustrated a gray area in between shared and configural constructs using an example of leadership research, in which there is debate on whether perceptions of a team leader can be shared among team members. Some scholars have suggested that a leader is likely to treat his or her subordinates in a similar way, whereas other researchers have countered that team leaders are likely to adjust their behavior to the specific team member they are interacting with (e.g., Henderson et al., 2009). For example, the team leader may be more friendly to more productive team members or more considerate to newer team members. In the latter cases, although the referent in an item may be the team leader, the individual perceptions may not truly tap into exactly the same cluster-level attribute. Similar processes may play a role when students report on behavior of teachers; the teacher may show different behavior to different students by adapting to their individual needs.

Using the item referent seems a natural way of defining constructs as being either shared or configural. However, in practice, one may encounter scales that consist of items that vary with respect to the target of measurement. For example, the Engaging Teaching scale from the Trends in International Mathematics and Science Study (TIMSS) 2015 data set consists of six items, of which five refer to the teacher (e.g., "My teacher is easy to understand") and one refers to the individual: "I know what my teacher expects." This scale was used as representing the measurement of a shared construct in Stapleton and Johnson (2019).

Such examples illustrate that it may not always be evident whether a construct should be viewed as shared or configural. Lüdtke et al. (2008) indicated that although their research was focused on shared and configural constructs (respectively, formative and reflective constructs in their terminology), these two types of constructs may actually represent opposite ends of a continuum. In practice, with data on multiple items rated by multiple cluster members, it is quite likely that constructs should be placed somewhere in the middle of the continuum from shared to configural. That is, for shared constructs, the individual perceptions of the cluster-level property will likely vary across cluster members. Even for an item such as "My school is fun to go to," a response provided by a student in a school will reflect that student's perception of the school. An item in which the referent is the individual, such as "I have fun at school," is unlikely to tap into a very different construct. The empirical effects of changing the item referent could be evaluated by gathering data based on two versions of items, one referring to the individual and one referring to the cluster (Keyton, 1991; Kirkman et al., 2001; Van Mierlo et al., 2009).

For configural constructs, in which the aggregate of the individual attributes differs across clusters, one may also argue that the actual cause of the cluster-level variance in the individual attributes is some cluster-level property. For example, one might be measuring students' mathematical achievement (clearly an individual-level construct) and find significant variance in the average mathematical achievement across classrooms. Possible causes of these differences can be classes having teachers who use different methods or have varying experience. If the cluster-level factor indirectly reflects such cluster-level properties, one would place the construct more to the middle than purely configural constructs on the continuum.

## *Suitability of using intraclass correlations to classify constructs*

In practice, researchers sometimes use statistical criteria such as values of intraclass correlations (ICCs; Bliese, 2000) to decide whether a construct can be considered shared. We do not think this is always appropriate. If a scale is designed to target a cluster-level construct, researchers may indeed hope that the majority of variation in the observed individual responses may exist at the cluster level rather than the individual level. In other words, one might hope that the IRR of the individuals rating the cluster construct is high. However, the ICC can also be relatively high for a configural construct. Consider, for example, students' socioeconomic status (SES). School populations can be very segregated on SES, leading to a large amount of variance at the school level (i.e., a large ICC). Thus, configural constructs may (however seldom) show higher proportions of cluster-level variance than shared constructs. Therefore, the ICC would not be an appropriate measure to classify constructs as shared or configural.

In the following sections, we illustrate that there is no need to specify different statistical models to represent a cluster-level construct as shared or configural and discuss how the ICCs can be interpreted instead. Similar to standard single-level factor analysis, the labeling and theoretical status of a latent variable should be based on theory, not on statistics. It is up to the researcher to determine how the common item variance at the cluster level can be interpreted and, therefore, what the theoretical status of the cluster-level constructs (measured through individual responses) is.

## Measurement Models for Cluster-Level Constructs With Individual-Level Data

Both shared and configural constructs are cluster-level constructs that are measured through individual-level responses and that require a measurement model. So far, we have discussed the shared and configural constructs separately. However, in this section, we show that from a statistical-modeling perspective, it is irrelevant whether a construct is thought to be shared or configural because the same statistical model can apply. Still, on the basis of the different definitions of configural, shared, and global constructs, several measurement models have been proposed to model cluster-level constructs with individual responses. In the next section, we therefore begin with a detailed explanation of the different proposed measurement models for cluster-level constructs that are operationalized by administering multiple items to the individuals in the cluster.

What all these methods have in common is that they account for the two-level structure in the data—originating from the individuals nested in clusters—using a multilevel extension of structural equation modeling (SEM). With multilevel modeling, an observation on variable $y$ for individual $i$ in cluster $j$ can be decomposed into a cluster component $y_j^{(B)} = \overline{y}_j$, representing the (latent) cluster mean on the variable, and an individual component, $y_{ij}^{(W)} = y_{ij} - \overline{y}_j$, representing an individual's deviation from the cluster mean. The cluster component $y_j^{(B)}$ reflects an unobserved cluster mean that takes the unreliability of the cluster mean (sampling error at the cluster level) into account (Lüdtke et al., 2008). Using the cluster means of the observed variables directly would not take sampling error into account because cluster means obtained from few individuals from actually large clusters would be regarded equally informative as cluster means obtained from a large proportion of individuals from large clusters.

SEM can account for measurement error in observations by incorporating a measurement model with multiple indicators $\mathbf{y}$ of a construct. In the multilevel extension of SEM (Muthén, 1989, 1994; Schmidt, 1969), each individual item score in the vector $\mathbf{y}_{ij}$ can be similarly decomposed into independent individual and cluster components. The covariance matrix of the individual components $y_{ij}^{(W)}$ is denoted $\Sigma^W$, and the covariance matrix of the cluster components $y_j^{(B)}$ is denoted $\Sigma^B$. That is, the total covariance matrix $\Sigma^T$ is decomposed into two orthogonal covariance matrices:

$$\Sigma^T = \Sigma^W + \Sigma^B.$$

With two-level SEM, one can fit factor models to explain the covariances at the individual and cluster levels. See Hox et al. (2017) for an introduction to two-level factor models.

### *Measurement models proposed for configural constructs*

Stapleton et al. (2016) explained that for configural constructs, the appropriate factor model is a two-level model with cross-level equality constraints on the factor loadings. The cross-level invariance of factor loadings is necessary to interpret the individual- and cluster-level common factors as reflecting the individual- and cluster-level components of the same construct (Asparouhov & Muthén, 2012; Hox et al., 2017; Kim et al., 2016; Lüdtke et al., 2011; Mehta & Neale, 2005; Muthén, 1990; Rabe-Hesketh et al., 2004; Stapleton et al., 2016). Cross-level invariance of factor loadings is implied by invariance of factor loadings across clusters (Jak et al., 2013). Hence, when metric invariance across clusters would be violated, then one would not be able to make valid comparisons across clusters because a 1-unit change in the (component of a) factor would not be linked with the same expected change in the (components of) indicators across clusters (or levels). Suppose that students' math achievement was measured with five tests completed by students in several classes, and the five tests are indicative of one common factor, "math achievement." One can make valid comparisons on math achievement across classes only when the factor loadings are invariant across classes, implying equal factor loadings across clusters in the two-level factor model. With cross-level invariance of factor loadings, the individual-level factor scores can be interpreted as individual deviations from the average math achievement in a class, and the cluster-level factor scores represent the average math achievement of each class.[1]

The doubly latent model proposed by Marsh et al. (2009) is equivalent to the measurement model proposed by Stapleton et al. (2016) depicted in Figure 2. The name "doubly latent" comes from the fact that the doubly latent model uses latent aggregation to take sampling error into account and uses latent variables with multiple

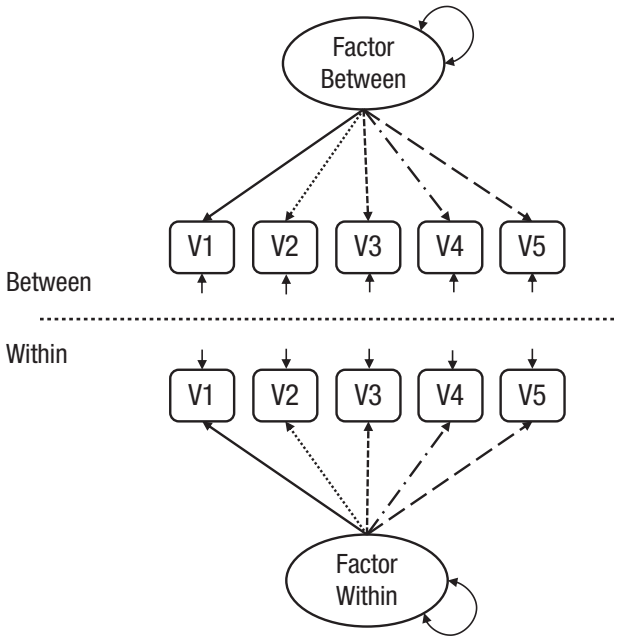


**Fig. 2.** Graphical illustration of the doubly latent model on five indicators. Factor loadings that share the same line type are constrained to be equal.

indicators to take measurement error into account. Marsh et al. did not explicitly state that cross-level invariance of factor loadings is necessary.[2] However, in a recent article on doubly latent multilevel procedures, Morin et al. (2022) did stress this requirement for a valid interpretation. Note that the cross-level invariance of factor loadings is testable, for example, by comparing the fit of a model with invariance constraints with a model without invariance constraints. If such a test indicates that cross-level invariance does not hold, this finding complicates the interpretation of the factors at the two levels. However, in practice, the assumption holds frequently, matching the finding that metric invariance across groups often holds (Boer et al., 2018; De Roover, 2021).

## *Measurement models proposed for shared constructs*

A large source of confusion surrounding the statistical modeling of shared constructs is that there seems to be a mismatch between how some authors define a shared construct in theory (i.e., as global constructs discussed earlier) and what is actually encountered in empirical research (i.e., some level of shared individual perceptions of cluster-level constructs). Imposing the statistical properties of global constructs on the measurement of what actually are shared constructs has led to the development of some questionable measurement models.

Marsh et al. (2012) used the doubly latent model for both shared (climate) and configural (contextual) constructs, which we also advocate. They used an identical statistical model with different interpretations about the type of construct at the cluster level, depending on the item wording. If the referent of items is the cluster (e.g., class or teacher), it represents a shared (climate) construct, and if the referent of the item is the individual (e.g., student), it represents a configural (contextual) construct. For shared constructs, the between-level construct represents the shared perceptions of the individuals in the cluster, and the individual-level construct represents individual deviations from the average perception in the cluster. Morin et al. (2022) and Jak (2019) noted that for shared constructs, cross-level invariance on the factor loadings in the doubly latent model should be imposed.

Initially, Stapleton et al. (2016) proposed fitting a saturated structure at the individual level for shared constructs because modeling the structure of individual-level responses should not be of interest for a cluster-level phenomenon. We argue that although this idea seems intuitive, such a model would indirectly accommodate violations of metric invariance across clusters, meaning that if a construct were measured in each cluster, its interpretation might differ across clusters. If this is the case, the cluster-level common factor becomes uninterpretable. In later work, Stapleton and Johnson (2019) abandoned the saturated model and advocated the "simultaneous shared-and-configural model" for shared constructs, originally proposed in Stapleton et al. Figure 3 shows an example of such a model with five items.

The simultaneous shared-and-configural model consists of one configural construct (with an individual- and cluster-level component[3]) and an additional orthogonal construct that exists only at the cluster level, intended to represent the truly shared construct. The configural construct is viewed as a nuisance construct, representing individual responses stemming from a construct that is unrelated to the shared construct and that may differ systematically across clusters. Stapleton and Johnson (2019) used the example of measuring neighborhood safety with items administered to residents in different neighborhoods. They stated that for such an example, an individual's responses to these items could reflect two dimensions: actual objective safety of the neighborhood (the target shared construct) and individual's tolerance for unsafe conditions (the nuisance configural construct). Note that a construct such as "actual objective safety" would be considered a global construct by Klein and Kozlowski (2000), so no measurement model would be needed. In accordance, Stapleton and Johnson stated that "in theory, a good measure of neighborhood safety would elicit the same response from all residents within

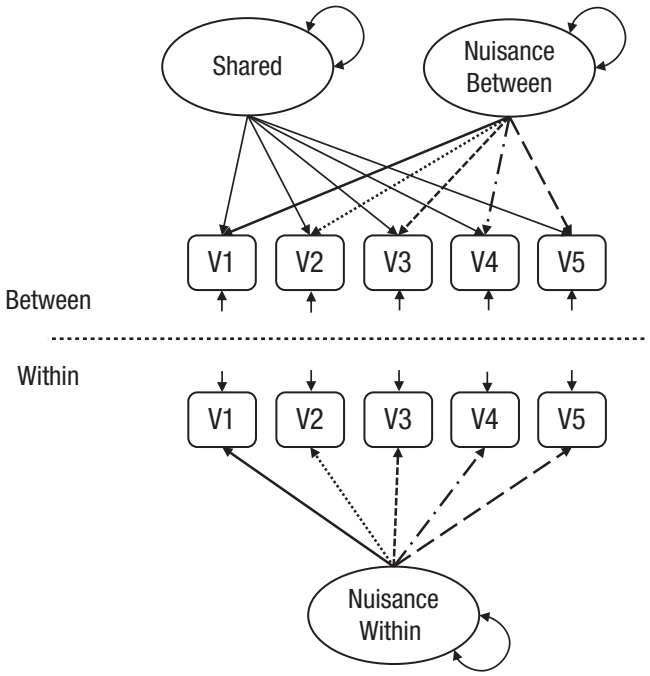


**Fig. 3.** Graphical representation of a simultaneous shared-and-configural model on five indicators. Factor loadings that share the same line type are constrained to be equal.

the same neighborhood" (p. 314), or in other words, a perfectly reliable measure of a shared construct would constitute a (truly shared) global construct.

We see multiple problems with the simultaneous shared-and-configural model. First, for objectively measured global properties, it would not make sense to operationalize them using ratings of all individuals in a cluster. One observation would be sufficient, yielding a manifest cluster-level variable. Second, objectively measured global properties will not be operationalized by several indicators that reflect a latent construct. Objective properties, such as the actual crime rate in a neighborhood, can be operationalized with a single variable or potentially as a formative construct using a combination of objective measures (Bollen & Bauldry, 2011). Third, even if the shared factor would indeed capture an objective property, there is no reason to assume that the within-clusters differences would reflect a single construct (e.g., "acquiescence" in Stapleton and Johnson's, 2019, illustrative example). Nuisance constructs need not be one-dimensional, given that each item could be affected by different, uncorrelated nuisance factors. Finally, modeling a shared construct as an additional between-level factor—orthogonal to a configural construct—does not necessarily extract the variance because of an objective, global construct from individual-level indicators. The shared construct might instead simply capture other noise or unmodeled multidimensionality, and its interpretation would be difficult. This seems to be related to why Stapleton and Johnson (2019) suggested the alternative approach of "us[ing] additional items, measured at the cluster level and not at the individual level" (p. 325).

In our view, when using multiple items and the responses of residents to measure neighborhood safety, one would be measuring individuals' perceptions of the neighborhood safety rather than objective neighborhood safety. Nonetheless, objective properties of neighborhoods, such as the actual crime rates, will influence the residents' perceptions, leading to shared perceptions within neighborhoods. Therefore, the shared perceptions may indeed provide information about neighborhood properties. However, we contest the interpretation of the shared factor in the shared-and-configural model as a purely objective cluster-level property, nor would we interpret the aggregate of perceptions about neighborhood safety as purely subjective or as reflecting acquiescence.

We argue that the doubly latent models from Marsh et al. (2012) are less conceptually problematic than Stapleton and Johnson's (2019) shared-and-configural model. Even if an item refers to the cluster, individuals' answers still reflect their perception of the cluster construct. The distinction between shared and configural constructs is therefore a theoretical one (when the configural construct is operationalized by aggregating cluster means), not a statistical one. In other words, researchers can use the doubly latent model to model both configural and shared constructs.

## *Quantifying the proportion of cluster-level variance in constructs*

The doubly latent model is flexible in the distribution of factor variance across levels. The more agreement there is, the higher is the proportion of cluster-level factor variance, and so the stronger is the shared part of the construct. One could quantify the amount of shared-construct variance by calculating the ICC of the common factors (Mehta & Neale, 2005), referred to as "$ICC_L$" by Kush et al. (2021):

$$ICC_L = ICC(1) = \frac{\Psi_C}{\Psi_C + \Psi_I},$$

with $\Psi_C$ and $\Psi_I$ representing the cluster- and individual-level factor variances, respectively. This $ICC_L$ may be useful to see how much of the construct's variance, as operationalized by the common factor, is shared across the cluster members. Although it might be tempting to use such a statistic to try to determine whether a construct is shared or configural, we believe it would be more useful to interpret it as a sort of IRR coefficient (McGraw & Wong, 1996; Shrout & Fleiss, 1979), as might

**Table 2.** Instructional Skills Questionnaire (Knol et al., 2016) Items and Intraclass Correlations Measuring the Dimension "Stimulation"

| Instructional Skills Questionnaire item | Intraclass correlation |
|---|---|
| 1. The lecture is boring (R) | .232 |
| 2. The instructor enlivens the subject matter | .300 |
| 3. It is hard to stay focused on the lecture (R) | .241 |
| 4. The instructor interests you in the subject matter | .240 |

Note: (R) denotes contra-indicative items.

be applied to factor scores (if they could be observed). Still, it may be informative to see, for example, what part of students' perceptions of teachers is shared among students because very low agreement may mean that asking the students may not be the most informative way to measure the specific property of the teacher.

## Illustrations

We illustrate the modeling of cluster-level construct using two examples. The first example involves data on students' evaluations of lecturers. The second example uses data of individuals' perceptions of conflict within their organizational team. The modeling procedures involve reflecting on the target of the items, quantifying the amount of variation in the observed variables at the individual and cluster levels, establishing an appropriate measurement model, evaluating the doubly latent model, and quantifying the proportion of construct variance at the team level. These data are available on OSF (https://osf.io/mzba8) with R syntax to replicate the analyses in `lavaan` (Rosseel, 2012).

### *Illustration 1*

Knol et al. (2016) designed a questionnaire to measure the quality of university lectures, called the Instructional Skills Questionnaire. The target of measurement is the lecture, the quality of which is evaluated by asking for ratings from multiple students attending the same lecture. The data were obtained for 5,422 students and 73 lectures. The response rate was 90.5% (Knol et al., 2016). For our illustration, we use the four items designed to measure the dimension "Stimulation," which are shown in Table 2. The items were scores on a 7-point Likert scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). Two of the items are contra-indicative ("hard to stay focused" and "lecture is boring"), and the scores on these items were recoded before analysis. The items' ICCs ranged from .232 to .300, indicating that roughly 75% of the items' variance was attributable to differences between students' perceptions of the same lecture and that around 25% was attributable to differences across lectures. Looking at the item content, it is not surprising to find substantial variance at the individual level. For example, it is quite likely that the same lecture is perceived as more boring by one student than by another student. Maybe some students are better prepared than others, or maybe students differ in their personal interests. The item referent is the instructor or lecture for three of the items, whereas Item 3 has no clear referent ("It is hard to stay focused on the lecture"). Students could either be answering this item referring to their own experience or trying to make an estimate of how hard it is for students in general to stay focused. The item with the highest ICC (Item 2) seems to allude less to personal experiences than the other three items. That is, the item refers to the instructor "enlivening the subject matter" and not to boredom, focus, or interest at the side of the student.

All in all, the items seem to measure students' individual perceptions of a lecture. Part of these perceptions is expected to be shared among students, and those shared perceptions may be indicative for the quality of lectures. Knol et al. (2016) applied factor analysis to the between levels only, with a saturated within-level model, using the argument that only the between-level model is of interest. This is not in line with our interpretations involving differences in student perceptions of lecture quality at the within levels and the shared part of those perceptions at the between levels. We cannot interpret the between-level construct as reflecting shared perceptions without modeling the individual perceptions construct at the student level. We therefore apply the doubly latent model to evaluate the student-level (within-level) and lecture-level (between-level) models.

### *Analysis*

The first step of the analysis was to identify a reasonable measurement model. There are multiple ways to approach this. One option is to immediately fit the doubly latent model. Alternatively, one could first establish a measurement model at the within levels while specifying a saturated model at the between levels or start with finding a measurement model at the between levels while fitting a saturated structure to the within levels. We think that modeling the between levels with a saturated within part is not sensible because the interpretation of the between-level construct depends on the within-level construct. This implies that when a measurement model is established at the between-parts only (with a saturated within-level model), then if in a next step the same factor structure is applied at the within level, the interpretation of the between-level constructs will change. Our advice is therefore to either consider the two levels together or to establish the measurement

model at the within level with a saturated between-level model. The latter approach is also in line with Bryk and Raudenbush's (1992) two-phase approach in ordinary multilevel regression and with the stepwise-modeling approach of multilevel mediation effects of Preacher et al. (2010). We therefore fitted a two-level model with a saturated model at the cluster level and the theoretically expected model structure at the individual level. This enabled us to calculate fit indices specific to the within-level model (Ryu & West, 2009). Next, we fitted the doubly latent model with cross-level invariance constraints on the factor loadings. If this model fitted the data well, we tested whether the residual variances at the cluster level could be constrained at zero. Zero residual variances at the cluster level in a model with cross-level invariance on factor loadings reflects scalar invariance across clusters (Jak et al., 2013). We tested hypotheses about equivalence (loadings) and fixed parameters (residual variances) using $\alpha = .05$ as criterion for statistical significance.

Model fit was evaluated using the $\chi^2$ test of exact fit, which will be rejected when the $\chi^2$ value is statistically significant at $\alpha = .05$. Approximate fit was evaluated with the root mean square error of approximation (RMSEA; Steiger & Lind, 1980), the comparative fit index (CFI; Bentler, 1990), and the level-specific versions when applicable. RMSEA values smaller than .05 were interpreted as indicating close fit, and values smaller than .08 were considered satisfactory (Browne & Cudeck, 1992). CFI values over .95 were interpreted as indicating reasonably good fit (Hu & Bentler, 1999). In case of inacceptable model fit, we evaluated modification indices (Chou & Bentler, 1990) and added model parameters only when they seemed theoretically reasonable.

## Results

***Finding a measurement model.*** We first fitted a one-factor model to the within level with a saturated model at the between level. In this model, all misfit stems from the within level. For the calculation of the within-level CFI, we needed to fit the independence model to the within level with a saturated between level. This model, however, was unable to provide a converged solution. We therefore report only the level-specific RMSEA and the overall CFI instead of the level-specific CFI (i.e., we used the default independence model at both levels). The one-factor model did not fit the data adequately, $\chi^2(2) = 120.94$, $p < .05$, CFI = .98, $\text{RMSEA}_W = .105$, 90% confidence interval [CI] = [.089, .121]. The largest modification index pertained to adding a covariance between Item 1 and Item 3 or between Item 2 and Item 4. These are actually the pairs of negatively and positively formulated items, respectively. The unmodeled covariance likely shows a wording effect (Horan et al., 2003). We therefore added a covariance between the residuals of Item 1 and Item 3. This model fitted the data adequately. Exact fit was not rejected, $\chi^2(1) = 1.82$, $p = .177$, and approximate fit indices show values associated with good fit: CFI = 1.00, $\text{RMSEA}_W = .012$, 90% CI = [.000, .041]. The one-factor model with added residual covariance was considered the final measurement model.

***Fitting the doubly latent model.*** We fitted the doubly latent model with cross-level invariance based on the measurement model from the previous step. The fit of this model was good: $\chi^2(5) = 9.73$, $p = .083$, CFI = .999, $\text{RMSEA}^4 = .013$, 90% CI = [.000, .026]. For Item 4, the estimated residual variance at the between level was negative and not significantly different from zero ($\hat{\theta} = -.006$, $SE = .006$, $p = .326$). The between-level residual variance of the other three items was statistically significant according to the univariate Wald $z$ tests. We therefore fixed only the residual variance of Item 4 at the between level to zero. The overall fit of this final model was good: $\chi^2(6) = 10.59$, $p = .102$, CFI = .999, RMSEA = .012, 90% CI = [.000, .023]. A graphical display of the doubly latent model with parameter estimates is provided in Figure 4.

The common factor at the within level represents what is common in the stimulation items but differs across students attending the same lecture. We label the construct "student perception of stimulation." Because of the cross-level invariance of the factor loadings, the between-level construct can be interpreted as the "cluster average student perception of stimulation," which represents what is common to the four items and common to students attending the same lecture. In other words, we interpret the between-level common factor as shared perceptions of stimulation. The $\text{ICC}_L$ of the stimulation factor was .775 / (.775 + 1) = .437, indicating that around 44% of the variance in the latent variable student perceptions of stimulation is shared among students attending the same lecture.

We stop the analysis here, but the model could be extended by adding within- or between-level variables to explain part of the variance in the common factor. For example, if one has an operationalization of how prepared students were for the lecture, one could add this variable as a predictor of the common factor at the within level. Or one could add variables related to the instructor (e.g., gender, motivation, level of burnout) at the between level to explain differences in perceived stimulation across different instructors.

The nonzero residual variances at the between level represent cluster differences on the items that are not attributable to the common factor shared perceptions of stimulation. There must be other variables unique to Item 1, Item 2, and Item 4 that caused structural differences on these items at the lecture level. In the terminology of Jak et al. (2013, 2014), the three items show

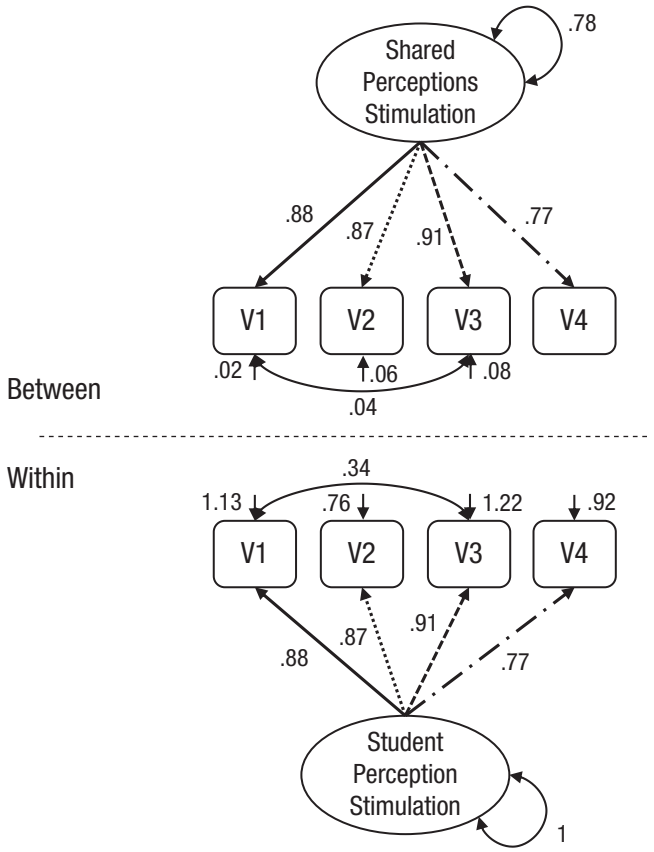


**Fig. 4.** Unstandardized parameter estimates of doubly latent model with cross-level invariance on four stimulation items. Factor loadings that share the same line type are constrained to be equal.

cluster bias. A follow-up analysis could also involve trying to explain the cluster bias using other between-levels variables. For example, the item "It is hard to stay focused on the lecture" could elicit different responses across lectures because of the timing of the lectures (Friday afternoon vs. Monday morning) or the physical circumstances, such as the temperature in the lecture hall; such factors have nothing to do with how stimulating the instructor made the lecture.

## *Illustration 2*

This illustration focuses on organizational teams instead of educational setting. We used data on eight items asking about the extent of relationship conflict and task conflict within organizational teams (Jehn, 1995). Example items are "How much friction is there between team members?" and "How often are there conflicting opinions about the work that has to be done?" The first four items reflect the construct "relationship conflict," and the last four items reflect the construct "task conflict." The items targeted the team, and the construct was not objectively quantifiable. Theoretically, the researchers thus assessed a shared construct, and a measurement model is required. Data were gathered from 228 employees who were employed in 113 teams. The included employees were selected by the team leaders. The response rate of employees was 88%. On average, the size of the teams was 15.66. The cluster size was two in 111 of the teams and three in two of the teams. The sampling ratio was therefore approximately .13. All items were scored on a 5-point Likert scale ranging from 1 (*none*) to 5 (*very much*). With these small cluster sizes and five response categories, there were multiple clusters in which the members completely agreed, resulting in zero variance within the cluster. For each item, complete agreement was observed in approximately 40 to 50 clusters. The proportion of cluster-level variation in the observed item scores was substantial and ranged from ICCs = .209 to .392 across items (see Table 3), indicating that despite the referent being the team, the largest part of the variance existed at the individual level for all items. We followed the same analysis procedure as in the previous illustration.

## *Results*

***Finding a measurement model.*** We started with fitting the two-factor model at the individual level, with a saturated cluster-level model. This way, all misfit arises from constraints at the individual level. As with the previous example, the within-level CFI could not be calculated because the independence model did not converge to a solution. We therefore provide the within-level RMSEA and the overall CFI. The two-factor model showed good fit to the data: $\chi^2(19) = 3.65$, $p = 1.00$, CFI = 1.00, $\text{RMSEA}_W$ = .00, 90% CI = [.00, .00]. The correlation between the two factors was substantial ($r = .72$, $p < .001$). We used the two-factor model as the final measurement model. The two individual-level factors model the within-teams differences in perceived relationship and task conflict or how members of the same team can have different perceptions of the conflicts in their team.

***Fitting the doubly latent model.*** Because the two-factor model at the individual level fitted the data well, we imposed the same structure at the cluster level. This two-level model with cross-level invariance on the factor loadings fitted the data well according to the approximate fit indices: $\chi^2(44) = 50.11$, $p = .024$, CFI = .99, RMSEA = .025, 90% CI = [.000, .053]. The cluster-level factors in this model represent the average perceived conflict within teams. The variance of the cluster-level factors represents differences in the team averages of perceived conflict. Residual variances at the cluster level reflect so-called cluster bias (Jak et al., 2013): team differences in item scores that cannot be fully attributed to team differences in the common factors. According to the univariate Wald $z$ tests with $\alpha = .05$, none

**Table 3.** The Conflict Items (Jehn, 1995) and Intraclass Correlations

| Item | Intraclass correlation |
|---|---|
| Relationship conflict | |
| 1. How much friction is there among members in your work unit? | .392 |
| 2. How much are personality conflicts evident in your work unit? | .359 |
| 3. How much tension is there among members in your work unit? | .384 |
| 4. How much emotional conflict is there among members in your work unit? | .298 |
| Task conflict | |
| 5. How often do people in your work unit disagree about opinions regarding the work being done? | .214 |
| 6. How frequently are there conflicts about ideas in your work unit? | .209 |
| 7. How much conflict about the work you do is there in your work unit? | .231 |
| 8. To what extent are there differences of opinion in your work unit? | .252 |

of the eight indicators has statistically significant residual variances at the cluster level. Consistent with scalar invariance across clusters (Jak et al., 2013; Jak & Jorgensen, 2017), fixing these residual variances to zero did not significantly deteriorate model fit, $\Delta\chi^2(8) = 12.74$, $p = .12$, and showed good overall fit, $\chi^2(52) = 62.85$, $p = .14$, CFI = .99, RMSEA = .030, 90% CI = [.000, .054]. In this model, the $ICC_L$ of the first factor was .81 / (.81 + 1) = .45, and the $ICC_L$ of the second factor was .49 / (.49 + 1) = .33, indicating that, respectively, 45% and 33% of the variance in the common factors exists at the cluster level. This finding suggests that there was more agreement among team members on relationship conflict than on task conflict. Figure 5 shows a graphical display of the doubly latent model with unstandardized parameter estimates. As in the previous example, the analysis could be extended by adding individual-level variables or team-level variables to the model. For example, one could test whether there are gender differences in the individual perceptions of conflict within teams by adding gender as a predictor of the within-level constructs. Or if one has an operationalization of the type of the leadership style employed by the manager, one could test whether differences in perceived conflicts between teams depend on leadership style.

## Discussion

In this study, we aimed to show that cluster-level constructs operationalized through individual-level responses hardly ever represent purely shared or purely configural constructs. We also argued that the distinction between the two types of cluster-level constructs is theoretical, not statistical. That is, one does not need different measurement models for constructs that are hypothesized to be shared or configural: The doubly latent model is suitable to both shared and configural constructs and anything ambiguously in between. In addition, we tried to disentangle the use of different terminology for the same type of constructs in the literature. In the following sections, we reflect on some issues related to modeling cluster-level constructs with individual-level responses.

### *Different types of cluster-level constructs from a theoretical perspective*

The main aim of this article was to dissuade the application of different models for theoretically shared versus configural constructs. We did not provide a solution to the question of how one can evaluate whether a cluster-level construct represents a theoretically shared or configural construct, but we advised against using statistical criteria (e.g., $ICC_L$) to do so. One could describe the models for shared and configural constructs as statistically equivalent but having different interpretations. Because the data are generally unable to provide direct information about the interpretation of cluster-level latent variables, it would instead be appropriate to focus on the meaning of cluster-level latent variables given the data (e.g., by carefully considering item content, as we demonstrated here). For comprehensive discussions of the conceptual status of cluster-level constructs, we refer to Klein and Kozlowski (2000), Chan (1998, 2019), and Morgeson and Hofmann (1999).

### *Estimation and specification issues in the doubly latent model*

Estimation issues are a common problem when modeling latent variables at multiple levels (e.g., Li & Beretvas, 2013; Lüdtke et al., 2011). Applying cross-level invariance on factor loadings, which is needed for interpretable common factors at both levels, already improves the likelihood of obtaining a converged solution using either frequentist or Bayesian estimation methods (Depaoli & Clifton, 2015; González-Romá & Hernández, 2017; Jak, 2019), even when cross-level invariance is not exactly true in the population (Kim & Cao, 2015). However, when there is a large difference in population factor loadings across levels, the doubly latent model with cross-level invariance is clearly misspecified. Researchers could then relax the equality constraint on the factor loadings for certain items, thereby allowing partial cross-level invariance of factor loadings (see e.g., Spilt et al., 2012). In

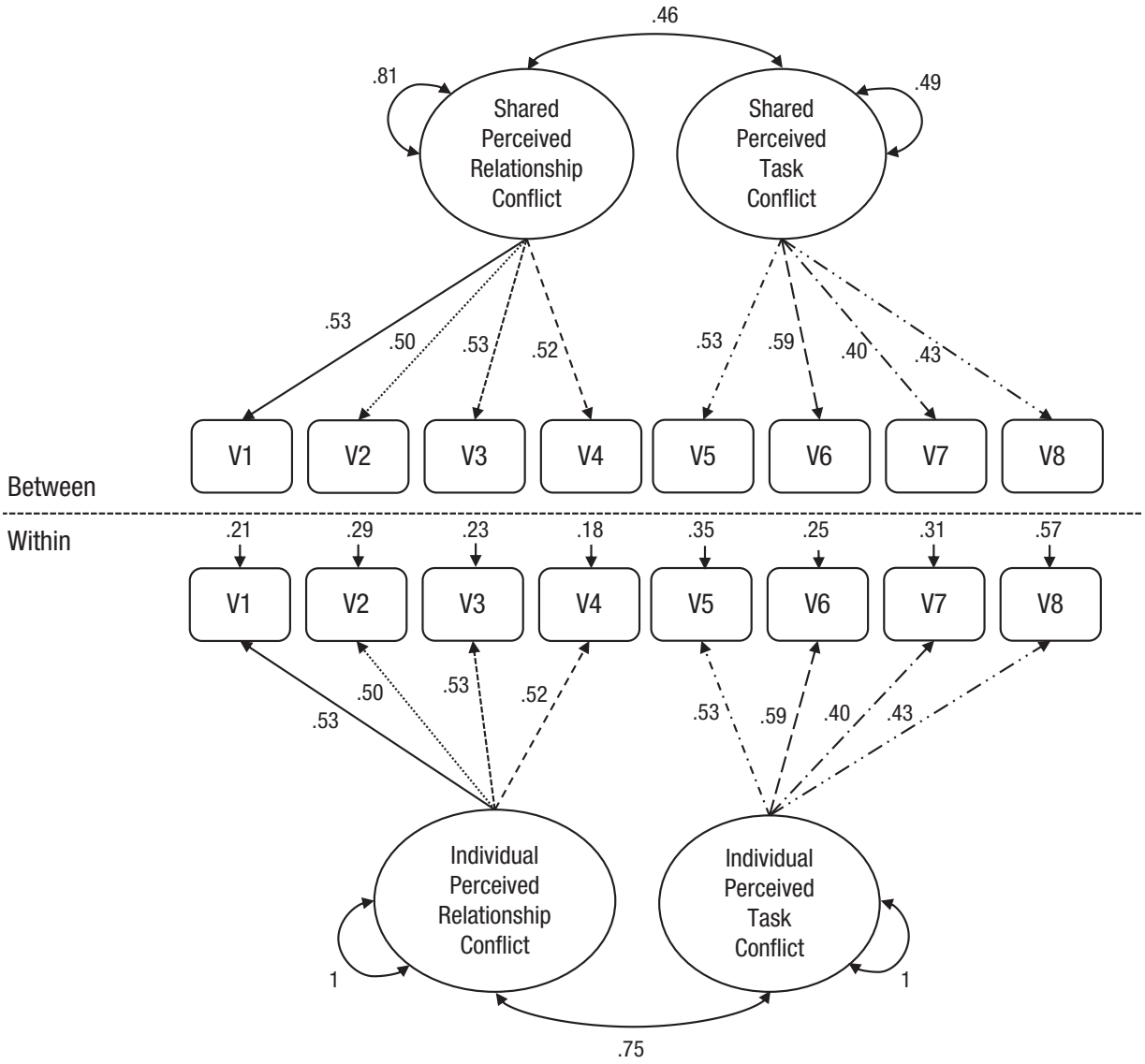


**Fig. 5.** Unstandardized parameter estimates of the doubly latent model with cross-level invariance on eight conflict items. Factor loadings that share the same line type are constrained to be equal.

situations in which even partial cross-level invariance does not hold, the decomposition of the construct into within-clusters and between-clusters components is dubious because the interpretation of the between-level construct becomes problematic (i.e., it does not necessarily represent the aggregate of something interpretable at the within level). Our advice would then be to stop the analysis and collect new data using a different measurement instrument.

A large source of estimation issues is related to the cluster-level residual variances. Specifically, convergence problems and negative variance estimates can be expected when population cluster-level residual variances are zero. Some software programs do not allow negative residual variances (Mplus; Muthén & Muthén, 1998–2017), but others do (`lavaan`; Rosseel, 2012). A recent evaluation of convergence issues in these two programs was conducted using data simulated from the configural (or doubly latent) model and from the simultaneous shared-and-configural model, either with zero or nonzero residual variances at the between level (Jak et al., 2021). It showed that convergence rates could be very different across algorithms, depending on whether the cluster-level residual variances were zero in the population or in the fitted model. For all conditions, `lavaan` either converged more often than Mplus, or both packages converged in 100% of samples. Mplus never converged in conditions in which cluster-level

residual variances were freely estimated while they were zero in the population. Rejection rates of the normal-theory $\chi^2$ test statistic were as expected and identical across packages, whereas rejection rates of the scaled test statistic were seriously inflated in several conditions. This inflation led the authors of Mplus to propose a correction on the robust chi-square statistic (Asparouhov & Muthén, 2021), implemented in Mplus Version 8.7.

## *Influence of the sampling ratio of individuals from clusters*

Two-level data often arise from a two-stage sampling design. In a two-stage design, researchers first sample the clusters and then sample individuals from clusters. One can imagine that increasing the number of individuals being sampled from each cluster leads to more reliable estimate of the cluster means. For example, if each cluster represents a school class with 25 children and one is interested in the average mathematical achievement in each school class, one would get better estimates when sampling 15 children from each class (sampling ratio of 15 / 25 = .60) than when sampling five children from each school class (sampling ratio of 5 / 25 = .20). Lüdtke et al. (2008) found that in the so-called multilevel latent covariate model, estimates of contextual effects suffered in situations with a low sampling ratio and a small number of individuals per cluster but were appropriately estimated in conditions with either large sampling ratios or large numbers of individuals per cluster. Guo et al. (2021) proposed and evaluated a finite population correction that leads to increased performance with medium-size sampling ratios. These results were obtained with a model that aggregates over the observed indicators instead of specifying a latent variable. A simulation study by Kush et al. (2021) focusing on the doubly latent model showed that lower sampling ratios have negative effects on parameter estimation. Specifically, the authors found bias in estimates of factor loadings and standard errors, although the size of the bias was considered negligible. In our first empirical illustration, the data came from 5,422 students nested in 73 lectures, so the cluster size was rather large (approximately 74). In this study, all students attending a lecture were invited to participate in the study, so the sampling ratio was 100%. In the second example, the cluster sizes were very small (teams of two or three employees), and only 13% of the participants were selected to participate in the study. With clusters as small as two or three and items that are scored on Likert-type scales, it is quite likely to find perfect agreement on items in some clusters. As a result, the cluster-level variance may be relatively large. It is important to realize that this might be the result of limited response options and a small cluster size in addition to cluster-level variability in the construct of interest.

It has been stated that the sampling ratio could be an important aspect when modeling configural constructs, but not for shared constructs, because for those constructs, the individual responses should (theoretically) be seen as exchangeable. Because we do not believe that in real research the individual responses about shared constructs can ever be seen as exchangeable (because then one would be evaluating a global construct rather than shared construct, so there would be no need to ask multiple individuals in a cluster), we argue that the sampling ratio could likewise be relevant for constructs that use individual perceptions of cluster-level properties (i.e., shared constructs). In other words, because the distinction between shared and configural constructs is potentially of theoretical value but not relevant to the statistical modeling of both types of construct, the sampling ratio would play a similar role for both types of theoretical constructs.

## *The emergence of shared constructs*

Shared constructs reflect agreement across group members, and it may be interesting to reflect on the development of the agreement among group members. One may expect that in newly formed groups, there may be less agreement than in longer existing groups. The process of group members increasing their agreement or similarity over time is referred to as "emergence" in the literature (Dansereau et al., 1999). Lang and Bliese (2018) developed the consensus-emergence model (CEM), which is essentially a three-level model of time points nested in individuals nested in organizations, specifying an exponential variance function on the within-groups variances. This function takes into account the expected decrease in within-groups variances as the group members agree more over time. The CEM currently focusses on single-indicator measurements of constructs. An interesting avenue for future research would be to extend the CEM to model emergence in common factors over configural time.

## *Conclusion*

Different terms have been used to point to the same type of cluster-level constructs (e.g., "configural," "contextual," and "formative" constructs), while at the same time, identical terms have been used to refer to different cluster-level concepts (e.g., the shared construct as defined by Stapleton et al., 2016, vs. the shared construct as defined by Klein & Kozlowski, 2000). We provided an overview of terminology and disentangled how the different terms are used in the literature. On the basis of this overview,

we argued that researchers do not need different models for theoretically shared versus configural constructs. The doubly latent model (Marsh et al., 2009) with cross-level invariance is the appropriate model for both types of constructs, whereas the simultaneous shared-and-configural model (Stapleton & Johnson, 2019) is ill defined and has uncertain interpretation.

## Transparency

*Action Editor:* Rogier Kievit
*Editor:* David A. Sbarra
*Author Contribution(s)*

**Suzanne Jak:** Conceptualization; Formal analysis; Methodology; Writing – original draft.

**Terrence D. Jorgensen:** Conceptualization; Methodology; Writing – review & editing.

**Debby ten Hove:** Investigation; Methodology; Writing – review & editing.

**Barbara Nevicka:** Data curation; Investigation; Writing – review & editing.

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

*Funding*

This work was partly supported by the Dutch Research Council funding awarded to S. Jak (Project VI.Vidi.201.009) and T. D. Jorgensen (Project 016.Veni.195.457).

## ORCID iDs

Suzanne Jak https://orcid.org/0000-0002-2223-5594
Terrence D. Jorgensen https://orcid.org/0000-0001-5111-6773
Debby ten Hove https://orcid.org/0000-0002-1335-4452

## Acknowledgments

Part of this research was presented at the Multilevel Conference in Utrecht, Netherlands, in 2022 and the 6th International NEPS conference, December 2021, Bamberg, Germany, and at the International Meeting of the Psychometric Society, 2022, Bologna, Italy.

## Notes

1. With cross-level invariance of factor loadings, the factor can be scaled by fixing the factor variance at 1 at one of the levels and freely estimating the factor variance at the other level. Alternatively, the factor variance can be estimated at both levels, under the constraint that they sum to 1; in this case, the between-levels variance equals the ICC of the factor scores.
2. Marsh et al. (2009) did apply the equality constraint in their presented analysis examples.
3. Stapleton and Johnson (2019) fixed the distribution of the configural factor's variance over levels, but this restriction is actually not needed (see Jak et al., 2021).
4. Level-specific fit indices are not meaningful for doubly latent models because it is impossible to combine cross-level invariance of factor loadings with fitting a saturated model at one of the levels. The reported RMSEA is therefore the overall RMSEA.

## References

Asparouhov, T., & Muthen, B. (2012). *Multiple group multilevel analysis* (Mplus Web Notes No. 16). http://statmodel.com/examples/webnotes/webnote16.pdf

Asparouhov, T., & Muthén, B. (2021). Robust chi-square in extreme and boundary conditions: Comments on Jak et al. (2021). *Psych*, *3*(3), 542–551. https://doi.org/10.3390/psych3030035

Bardach, L., Yanagida, T., & Lüftenegger, M. (2020). Studying classroom climate effects in the context of multi-level structural equation modelling: An application-focused theoretical discussion and empirical demonstration. *International Journal of Research & Method in Education*, *43*(4), 348–363. https://doi.org/10.1080/1743727X.2020.1791071

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

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). Jossey-Bass.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology*, *49*(5), 713–734. https://doi.org/10.1177/0022022117749042

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*(2), 230–258.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical models: Applications and data analysis methods*. Sage.

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. https://doi.org/10.1037/0021-9010.83.2.234

Chan, D. (2019). Team-level constructs. *Annual Review of Organizational Psychology and Organizational Behavior*, *6*, 325–348. https://doi.org/10.1146/annurev-orgpsych-012218-015117

Chou, C. P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research*, *25*(1), 115–136. https://doi.org/10.1207/s15327906mbr2501_13

Dansereau, F., Yammarino, F. J., & Kohles, J. C. (1999). Multiple levels of analysis from a longitudinal perspective: Some

implications for theory building. *Academy of Management Review*, *24*(2), 346–357.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling*, *22*, 327–351. https://doi.org/10.1080/10705511.2014.937849

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, *17*, 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

González-Romá, V., & Hernández, A. (2017). Multilevel modeling: Research-based lessons for substantive researchers. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 183–210. https://doi.org/10.1146/annurev-orgpsych-041015-062407

Guo, S., Houang, R. T., & Schmidt, W. H. (2021). The decomposition of between and within effects in contextual models. *Frontiers in Psychology*, *12*, 541803.

Henderson, D. J., Liden, R. C., Glibkowski, B. C., & Chaudhry, A. (2009). LMX differentiation: A multilevel review and examination of its antecedents and outcomes. *The Leadership Quarterly*, *20*(4), 517–534. https://doi.org/10.1016/j.leaqua.2009.04.003

Horan, P. M., DiStefano, C., & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, *10*(3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*, *1*, 179–188.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge. https://doi.org/10.4324/9781315650982

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

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling*, *26*, 607–622. https://doi.org/10.1080/10705511.2018.1534205

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology*, *8*, Article 1640. https://doi.org/10.3389/fpsyg.2017.01640

Jak, S., Jorgensen, T. D., & Rosseel, Y. (2021). Evaluating cluster-level factor models with lavaan and Mplus. *Psych*, *3*(2), 134–152. https://doi.org/10.3390/psych3020012

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*, 265–282. https://doi.org/10.1080/10705511.2013.769392

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

Jehn, K. A. (1995). A multimethod examination of the benefits and detriments of intragroup conflict. *Administrative Science Quarterly*, *40*(2), 256–282. https://doi.org/10.2307/2393638

Keyton, J. (1991). Evaluating individual group member satisfaction as a situational variable. *Small Group Research*, *22*(2), 200–219. https://doi.org/10.1177/1046496491222004

Kim, E. S., & Cao, C. (2015). Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. *Multivariate Behavioral Research*, *50*, 436–456. https://doi.org/10.1080/00273171.2015.1021447

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

Kirkman, B. L., Tesluk, P. E., & Rosen, B. (2001). Assessing the incremental validity of team consensus ratings over aggregation of individual-level data in predicting team effectiveness. *Personnel Psychology*, *54*(3), 645–667. https://doi.org/10.1111/j.1744-6570.2001.tb00226.x

Klein, K. J., & Kozlowski, S. W. (2000). *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. Jossey-Bass.

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

Kush, J. M., Konold, T. R., & Bradshaw, C. P. (2021). The sampling ratio in multilevel structural equation models: Considerations to inform study design. *Educational and Psychological Measurement*, *82*(3), 409–443. https://doi.org/10.1177/00131644211020112

Lange, J. W. B., & Bliese, P. D. (2019). A temporal perspective on emergence: Using three-level mixed-effects models to track consensus emergence in groups. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 519–540). American Psychological Association. https://doi.org/10.1037/0000115-023

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling*, *20*(2), 241–264. https://doi.org/10.1080/10705511.2013.769391

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. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent

covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

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

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*(6), 764–802. https://doi.org/10.1080/00273170903333665

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*(2), 249–265. https://doi.org/10.5465/amr.1999.1893935

Morin, A. J., Blais, A. R., & Chénard-Poirier, L. A. (2022). Doubly latent multilevel procedures for organizational assessment and prediction. *Journal of Business and Psychology*, *37*, 47–72. https://doi.org/10.1007/s10869-021-09736-5

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Muthén, B. (1990). *Mean and covariance structure analysis of hierarchical data* (UCLA Statistics Series #62). UCLA.

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

Muthén, L. K. & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Preacher, K., Zyphur, M., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233. https://doi.org/10.1037/a0020141

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika*, *69*, 167–190.

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

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

Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model* [Unpublished doctoral dissertation, Department of Education, University of Chicago].

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Spilt, J. L., Koomen, H. M., & Jak, S. (2012). Are boys better off with male and girls with female teachers? A multilevel investigation of measurement invariance and gender match in teacher–student relationship quality. *Journal of School Psychology*, *50*(3), 363–378. https://doi.org/10.1016/j.jsp.2011.12.002

Stapleton, L. M., & Johnson, T. L. (2019). Models to examine the validity of cluster-level factor structure using individual-level data. *Advances in Methods and Practices in Psychological Science*, *2*(3), 312–329. https://doi.org/10.1177/2515245919855039

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

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of factors* [Paper presentation]. Annual Spring Meeting of the Psychometric Society, Iowa City, IA, United States.

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, *12*(2), 368–392. https://doi.org/10.1177/1094428107309322