MULTIVARIATE BEHAVIORAL RESEARCH
2026, VOL. 61, NO. 5, 788–805

OPEN ACCESS

Single-Level Bifactor Models as Implicit Multilevel Factor Models Without a Bifactor Structure

Christian L. L. Strauss and Kristopher J. Preacher

Department of Psychology and Human Development, Vanderbilt University Peabody College, Nashville, TN, USA

ABSTRACT

It is well-known that bifactor structures are over-represented as preferred solutions in measurement modeling. This study explores the extent to which unmodeled clustering of observations in larger social or organization units (e.g., students clustered in schools) offers a partial explanation for this phenomenon. We investigate the overlap between bifactor confirmatory factor models and multilevel confirmatory factor analysis fit to identically structured data formats. Structural symmetries between these models are identified, leading to a series of postulates regarding expected differences across modeling frameworks. Next, through simulation and empirical data analysis, we demonstrate that bifactor solutions can emerge as artifacts of conflated level-1 and level-2 effects when clustering is ignored, causing invalid interpretations of factors. Specifically, results suggest that when bifactor models are fit to clustered data with one level-2 factor and multiple level-1 factors, general factor loadings are typically inflated, leading to greater support for the misspecified bifactor solution. We encourage researchers to consider multilevel measurement models as alternative explanations for bifactor solutions, so factors are accurately interpreted at the correct level of analysis.

KEYWORDS

Bifactor model; confirmatory factor analysis; multilevel confirmatory factor analysis; measurement models

Introduction

In the social and behavioral sciences, data collected for the purpose of instrument and scale development often involve item responses obtained from individuals clustered in larger social and/or organizational units. For example, a 2021 study exploring potential overlap in burnout and depression, involving a sample of educational staff members recruited from schools further nested in districts, found support for construct overlap between depression and exhaustion using an exploratory bifactor model (Verkuilen et al., 2021). In addition, a 2018 study aimed at validating a patient-physician shared decision-making scale sampled patients clustered in a smaller sample of medical providers. This study also found support for an underlying bifactor structure using both exploratory and confirmatory factor analysis (Calderon et al., 2018). Notably, in these examples and others from recent literature (Bae et al., 2020; Bae & DeBusk-Lane, 2019; Bartroli et al., 2022; Bianchi & Schonfeld, 2021; Calderon et al., 2018; Calvete et al., 2023; Caspi et al., 2024; Cuesta-Vargas et al., 2018; Dierendonck et al., 2020; Disabato et al., 2019; Maghsoodi et al., 2023; Ondé et al., 2022; Perera et al., 2018; Perreira et al., 2018; Quattrone et al., 2019, 2019; Savahl et al., 2023; Scherer et al., 2017; Shuck et al., 2019; Tiego et al., 2020; Torres-Vallejos et al., 2021; Verkuilen et al., 2021), single-level exploratory and/or confirmatory factor analyses were used to uncover underlying measurement structures for the purpose of scale development and construct clarity. Further, data were likely characterized by non-independent observations due to naturally occurring, or design-induced, clustering (e.g., sampling from schools, countries, vocational organizations)[1]. Lastly, all cited analyses concluded that a bifactor structure resulted in an optimally or well-fitting solution.

To summarize, single-level factor models are often used to model measurement structures in data that are likely clustered within measured or unmeasured hierarchical groups. In addition, past literature has

[1]Cited articles did not report item-level ICCs. Therefore, we can only conjecture that conflating level-1 and level-2 effects in a single-level model would result in an uninterpretable solution.

© 2026 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

clearly stated that bifactor structures often fit well, and may be over-represented as the optimal data generating process in scale development literature (Eid et al., 2017; Murray & Johnson, 2013; Reise, 2012; Reise et al., 2023; Rodriguez et al., 2016). The purpose of this paper is to discuss how multilevel measurement models can *masquerade* (to employ terminology from Bauer & Cai, 2009; Belzak & Bauer, 2019; Lubinski & Humphreys, 1990) as bifactor models under certain conditions. We will demonstrate that an applied researcher is likely to find empirical support for a bifactor model when clustering of individuals in social and/or organizational units is nontrivial but not decomposed. As a consequence, this researcher might then interpret cluster-level factors at lower levels of analysis, leading to invalid interpretations of the true underlying measurement structure.

Before returning to unique challenges in bifactor modeling, we begin by discussing the potential limitations of employing measurement models that do not decompose level-1 and level-2 effects in clustered data structures broadly. It is well known that clustered data violate the independence assumption, leading to biased standard errors in regression models and other statistical methods. Despite this, if data are collected solely for measurement modeling purposes (e.g., factor analysis), there is typically less reliance on standard errors, test-statistics, and p-values. Thus, violations of independence assumptions may be seemingly less consequential. Perhaps, for this reason, multilevel approaches to measurement modeling are applied less frequently than their single-level counterparts, even when data are not independent.

There are, however, additional consequences of applying single-level modeling frameworks to clustered data structures. In the context of scale development, when single-level models such as confirmatory factor analysis (CFA; e.g., Brown, 2006) are applied to multilevel data where observed variables have nontrivial intraclass correlation coefficients (ICCs), resulting solutions represent a conflation of level-1 and level-2 effects (Asparouhov & Muthn, 2006; Enders & Tofighi, 2007; Hedeker & Gibbons, 2006; Kreft et al., 1995; Lüdtke et al., 2008; Mancl et al., 2000; Miller & Burstein, 1981; Neuhaus & Kalbfleisch, 1998; Neuhaus & McCulloch, 2006; Preacher et al., 2010, 2016; Raudenbush & Bryk, 2002). For multilevel data, this can manifest in solutions that do not accurately represent the true measurement structure underlying either level-1 or level-2. As a solution, multilevel measurement modeling techniques may be employed, wherein latent variable models are separately and simultaneously fit to decomposed level-1 and level-2 covariance matrices of manifest variables (Hox, 1993; Hox et al., 2018; Kamata, 2001; Muthén, 1994). Importantly, this allows for the systematic evaluation of unconflated or unique level-1 and level-2 factor structures. Further, the extent to which level-1 and level-2 factor models differ, or cross-level measurement non-invariance, can be empirically investigated (Jak, 2019; Jak et al., 2013, 2014; Jak & Jorgensen, 2017; Zyphur et al., 2008).

Interestingly, a 2016 review found that nearly one third of multilevel measurement models in the literature resulted in an optimally fitting solution with a different number of level-1 and level-2 factors (Kim et al., 2016). This is often referred to as cross-level configural non-invariance. Further, on average, level-2 models were shown to have fewer factors than level-1 models. These findings suggest that it is relatively common for level-1 and level-2 factor structures to be structurally and functionally non-invariant. In the presence of cross-level non-invariance, models that do not decompose level-1 and level-2 effects may not, and likely do not, accurately represent the true measurement structure at either level of analysis.

This work is particularly salient to augment the current state of methodological literature on both bifactor measurement models and multilevel measurement models. Specifically, it is well known that fit indices in factor analysis tend to overly prioritize bifactor structures, which has led to the critical development of novel models and measures to help analysts more effectively uncover true underlying factor structures (Eid et al., 2017; Reise, 2012; Reise et al., 2016; Rodriguez et al., 2016). The overlap between bifactor models and multilevel measurement models with fewer factors at level-2 may in-part explain the ubiquity of bifactor models and offer potential alternative modeling solutions that may better map onto theory. In addition, with the growth in application of multilevel modeling frameworks and continued encouragement to think critically about their use (McNeish et al., 2017; Stapleton, McNeish, et al., 2016; Stapleton, Yang, et al., 2016), it is imperative to thoroughly explore potential consequences of conflating level-1 and level-2 effects in a variety of applications.

We will begin with a brief review of traditional bifactor confirmatory factor models, followed by multilevel confirmatory factor analysis, to introduce a shared notation which will be used throughout. After this, we will introduce structural symmetries in manifest variable and model-implied covariance structure equations underlying both models. Next, we will explore the extent to which these symmetries do not imply equality in conditions commonly encountered in practice. We then further explore the sample

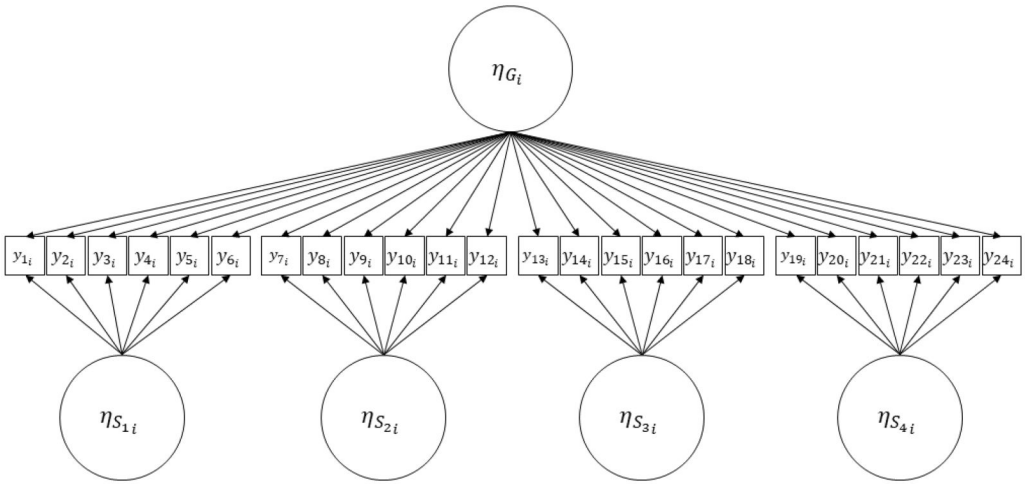

Figure 1. Path diagram of a bifactor model with one general factor and four specific factors. *Note*. Errors and mean structure are omitted for simplicity.

behavior of this inequality through a simulation and empirical demonstration.

Bifactor models

A bifactor model is a factor model consisting of at least one general factor and two or more specific factors[2]. More technically, this model assumes covariances among observed variables can be completely decomposed into a part due to a general factor and a part due to orthogonal specific factors. Consider the following equations underlying the CFA, assuming standardized latent variables for simplicity (i.e., the factor mean structure is omitted):

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \tag{1}$$

$$\boldsymbol{\eta}_i = \boldsymbol{\zeta}_i \tag{2}$$

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_{\epsilon} \tag{3}$$

where $\mathbf{y}_i$ is vector of manifest variable data for individual i, $\boldsymbol{\nu}$ is a vector of intercepts, $\boldsymbol{\Lambda}$ is matrix of factor loadings, $\boldsymbol{\epsilon}_i$ is a vector of unique factors for individual i with $\boldsymbol{\epsilon}_i \tilde{N}(\mathbf{0}, \boldsymbol{\Theta}_{\boldsymbol{\epsilon}})$, and $\boldsymbol{\zeta}_i$ is a vector of factor residuals with $\boldsymbol{\zeta}_i \tilde{N}(\mathbf{0}, \boldsymbol{\Psi})$. In the case of a bifactor model, one column vector of $\boldsymbol{\Lambda}$ contains freely estimated factor loadings for items on the general factor (wherein either one loading can be fixed to one for identification purposes, or the mean and variance of latent variables can be fixed to zero and one respectively, and all items are freely estimated), and the remaining column vectors contains loadings on the specific factors, with some loadings fixed to zero, as in a standard CFA. Further, off-diagonal elements of $\boldsymbol{\Psi}$ are fixed to zero for model identification purposes and to align with substantive implications of processes underlying manifest variable covariances. This aspect of the traditional bifactor model will be of great importance when discussing symmetries with multilevel factor models. Figure 1 displays a general path diagram for a bifactor model based on Equation (1).

Thus far, we have solely discussed single-level bifactor models[3], that is, factor models fit to a single and potentially conflated (when data are clustered) covariance matrix. Next, we introduce a multilevel approach to factor analysis, specifically multilevel confirmatory factor analysis (MCFA).

Multilevel confirmatory factor analysis

MCFA allows for the decomposition of a covariance matrix into constituent and orthogonal level-1 and level-2 components. That is, item responses are modeled additively with level-1 (within-cluster) and level-2 (between-cluster) parameters. Paralleling Equations (1)-(3), these can be expressed as:

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\Lambda}_B\boldsymbol{\eta}_j + \boldsymbol{\Lambda}_W\boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{B_j} + \boldsymbol{\epsilon}_{W_{ij}} \tag{4}$$

$$\boldsymbol{\eta}_j = \boldsymbol{\zeta}_{Bj} \tag{5}$$

[2]"Specific factors" are often referred to as "group factors" in the bifactor literature. We use "specific factor" here and throughout to avoid confusion with groups or clusters in multilevel modeling.

[3]We employ the term "single-level" bifactor models here and elsewhere to acknowledge that the source of clustering we are concerned with in this investigation stems from individuals clustered in social and/or organizational units (e.g., students in schools). This is not to be confused with the 2-level sampling procedure required for bifactor modeling as outlined in Eid et al. (2017). By "single-level" we mean that the model in question does not decompose clustering effects in the traditional sense, as discussed in multilevel literature (Asparouhov & Muthén, 2006; Enders & Tofighi, 2007; Hedeker & Gibbons, 2006; Kreft et al., 1995; Lüdtke et al., 2008; Mancl et al., 2000; Miller & Burstein, 1981; Neuhaus & Kalbfleisch, 1998; Neuhaus & McCulloch, 2006; Preacher et al., 2010, 2016; Raudenbush & Bryk, 2002).

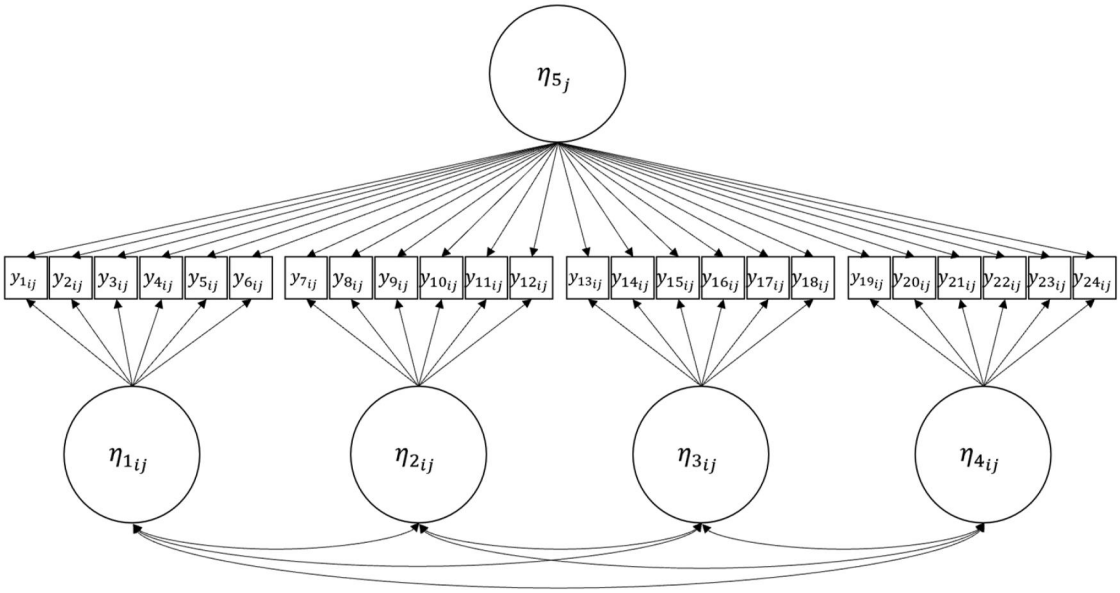

Figure 2. Path diagram of a multilevel confirmatory factor model with one level-2 factor and four level-1 factors. *Note*. Errors and mean structure are omitted for simplicity.

$$\boldsymbol{\eta}_{ij} = \boldsymbol{\zeta}_{W_{ij}} \tag{6}$$

$$\boldsymbol{\Sigma}_B(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_B \boldsymbol{\Psi}_B \boldsymbol{\Lambda}_B{}' + \boldsymbol{\Theta}_{\epsilon_{Bj}} \tag{7}$$

$$\boldsymbol{\Sigma}_W(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_W \boldsymbol{\Psi}_W \boldsymbol{\Lambda}_W{}' + \boldsymbol{\Theta}_{\epsilon_{Wij}} \tag{8}$$

Here $\mathbf{y}_{ij}$ refers to manifest variables for individual i in cluster j, $\boldsymbol{\mu}$ represents the fixed-effect associated with random intercepts given by $\boldsymbol{\mu}_j$ which represent latent cluster item means, $\boldsymbol{\Lambda}_W$ and $\boldsymbol{\Lambda}_B$ represent the fixed within- and between-factor loading matrices respectfully, $\boldsymbol{\epsilon}_{Wij}$ and $\boldsymbol{\epsilon}_{Bj}$ represent the level-1 and level-2 unique factor vectors, such that $\boldsymbol{\epsilon}_{Wij}\tilde{N}(\mathbf{0}, \boldsymbol{\Theta}_{\epsilon_W})$ and $\boldsymbol{\epsilon}_{Bj}\tilde{N}(\mathbf{0}, \boldsymbol{\Theta}_{\epsilon_B})$. In addition, level-1 factor residuals $\boldsymbol{\zeta}_{W_{ij}}$ are distributed as $\boldsymbol{\zeta}_{W_{ij}}\tilde{N}(\mathbf{0}, \boldsymbol{\Psi}_W)$ and level-2 factor residuals $\boldsymbol{\zeta}_{Bj}$, are distributed $\boldsymbol{\zeta}_{Bj}\tilde{N}(\mathbf{0}, \boldsymbol{\Psi}_B)$. For simplicity, we omit level-2 factor means and mean structures, but these can be incorporated into Equation (5). The mean structure is captured solely at level-2, therefore there is not a level-1 mean structure (Ryu, 2014).

Notably, the separate within- and between-cluster matrices imply matrix dimensions and elements can differ across level-1 and level-2. For the purpose of relating this model to the traditional bifactor model, $\boldsymbol{\eta}_j$ can be vector containing a single element and $\boldsymbol{\eta}_{ij}$ can be a larger vector. Further, separate within- and between-covariance structures imply elements of these matrices, including factor loadings, can differ across level-1 and level-2. Figure 2 displays a path diagram of an MCFA with one level-2 factor and four level-1 factors, based on Equation (4). Importantly, the elements of Equation (4) not involving error structures or mean structures imply an analogous path diagram to that of Figure 1. We include correlations among level-1 factors, or elements of $\boldsymbol{\eta}_{ij}$, as these are estimable in the MCFA.

While a variety of measurement structures can be specified and empirically evaluated using the MCFA framework, we will exclusively consider structures similar to what is depicted in Figure 2 for the purpose of comparison to traditional bifactor models, though findings may apply to models with more than one level-2 factor. Next, we further motivate the potential relationship between these models by noting a structural symmetry across model equations. This structural symmetry will be used to determine conditions for inequality of parameter estimates across these models, which will then inform a deeper exploration of parameter estimate behavior in subsequent simulation demonstrations.

Structural symmetry

Before presenting structural symmetries implied by Figures 1 and 2, we will clarify differences between what we will demonstrate and what past literature has proven regarding convergence between multilevel and bifactor-like factor models (e.g., Cai & Houts, 2021; Castro-Alvarez et al., 2022; Eid et al., 2017, 2017; Ulitzsch et al., 2017). Specifically, literature has established that many hierarchical and bifactor-like factor structures (e.g., latent curve models, latent state-trait models, and multitrait-multimethod models) can be reparameterized as multilevel measurement models. This typically involves first transforming data between

"long" format and "wide" format or a "wide" block structure. Then, a multilevel measurement model is estimated with a reduction in dimensionality at one or both levels of analysis. For example, Cai & Houts (2021) showed that a latent curve model with five factors used to denote Intercept, Slope, and three time-point-specific factors can be reparameterized as a three-factor multilevel model with an Intercept and Slope at level-2 and a single "time" factor at level-1. In addition, Ulitzsch et al. (2017) showed that a multi-trait-multimethod (MTMM) model with two methods (or raters) and two traits, leading to six total factors, can be reparameterized as a multilevel model with only four factors: two level-2 trait factors and two level-1 method (or rater) factors.

The symmetries we present are functionally different than these reparameterizations for two critical reasons. First, our symmetries hold in the absence of shifting data between "wide" and "long" format. Second, our symmetries do not change the dimensionality of either the bifactor model or the multilevel model without a bifactor structure, leading to equivalent dimensions across the symmetries. For example, Figures 1 and 2 show that a bifactor model with one general factor and four specific factors (five total factors) is observationally similar to a multilevel factor model with one level-2 factor and four level-1 factors (five total factors). This is important in application because, as our symmetries and subsequent simulation will demonstrate, an analyst working with data structured appropriately for a bifactor model (i.e., in "wide" format), can still be led astray by the masquerading phenomenon of the models in Figures 1 and 2 if level-1 and level-2 effects due to clustering of individuals in social/organizational units are not decomposed. That is, if the true data-generating mechanism is a multilevel measurement model without a bifactor structure, a single-level bifactor model is an inappropriate and invalid alternative model even though, as we will demonstrate, it is likely to be empirically supported. Thus, we refer to this phenomenon as a masquerade.

To begin, we assume Equations (1)-(3) represent a bifactor structure and Equations (4)-(8) represent a multilevel structure without a bifactor structure. For simplicity, we will consider a multilevel population generating structure with a single level-2 factor in the MCFA corresponding to a single general factor in the bifactor model, and multiple level-1 factors corresponding to multiple specific factors, though all subsequent symmetries expand to more than one level-2 or general factor. We start by re-expressing Equation (1) as:

$$\mathbf{y}_i = \boldsymbol{\nu} + \boldsymbol{\Lambda}_G \boldsymbol{\eta}_{G_i} + \boldsymbol{\Lambda}_S \boldsymbol{\eta}_{Si} + \boldsymbol{\epsilon}_i \tag{9}$$

Where $\boldsymbol{\Lambda}_G$ is a vector of loadings on the general factor, $\boldsymbol{\Lambda}_S$ is a matrix of loadings on specific factors, $\boldsymbol{\eta}_{Gi}$ is a scaler containing the true factor score on the general factor for person i, and $\boldsymbol{\eta}_{Si}$ is a vector of true factor scores for the specific factors for person i.

We can now define the following symmetries relating Equation (4) and Equation (9):

$$\boldsymbol{\mu} \approx \boldsymbol{\nu} \tag{10}$$

$$\boldsymbol{\Lambda}_B \approx \boldsymbol{\Lambda}_G \tag{11}$$

$$\boldsymbol{\eta}_j \approx \boldsymbol{\eta}_{G_i} \tag{12}$$

$$\boldsymbol{\Lambda}_W \approx \boldsymbol{\Lambda}_S \tag{13}$$

$$\boldsymbol{\eta}_{ij} \approx \boldsymbol{\eta}_{Si} \tag{14}$$

Where $\approx$ is used to denote symmetric elements of equations and not precise equality. That is, Symmetries (10)-(14) do not necessarily imply equivalency of individual elements of each corresponding matrix.

Next, we present specific details regarding the scenarios in which Symmetries (10), (11), and (14) should not lead to precise equivalencies. We begin this discussion by returning to analytics, which can be used to more clearly demonstrate the inequality between elements of matrices. The analytics, however, do not clarify sample manifestations of the inequality. As a solution, we explore the sample behavior of the inequality in a subsequent simulation demonstration.

Symmetry of intercepts

We begin with Symmetry (11). Recall, $\boldsymbol{\mu}_j$ is a random effect and therefore not directly estimated in the MCFA. Instead, fixed effects, $\boldsymbol{\mu}$, are estimated for each observed variable $\mathbf{y}_{ij}$. To further explicate, we can decompose Equation (4) into its constituent level-1 and level-2 components:

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\Lambda}_W \boldsymbol{\eta}_{ij} + \boldsymbol{\epsilon}_{Wij} \tag{15}$$

$$\boldsymbol{\mu}_j = \boldsymbol{\mu} + \boldsymbol{\Lambda}_B \boldsymbol{\eta}_j + \boldsymbol{\epsilon}_{Bj} \tag{16}$$

Without loss of generality, assuming the model is identified by setting latent variable means to 0, $\boldsymbol{\mu}$ is the expected score on the latent indicator $\boldsymbol{\mu}_j$ for a cluster with an average value of the latent variable(s) in $\boldsymbol{\eta}_j$. In contrast, according to Equation (9) $\boldsymbol{\nu}$ is the expected score on observed variable $\mathbf{y}_i$ for an individual who is average on both specific and general latent variables. Therefore, $\boldsymbol{\mu}$ is a cluster-level expected value

and $\boldsymbol{\nu}$ is an individual-level expected value. As population parameters, where sample sizes are irrelevant, these expected values should be equivalent. In practice, however, sample estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, namely $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\nu}}$ should be influenced by cluster size. Specifically, $\hat{\boldsymbol{\mu}}$ is an unweighted expected value across clusters, meaning clusters of different sizes equally contribute to the estimation of $\hat{\boldsymbol{\mu}}$. Following this logic, an unweighted expected value across clusters will only be equivalent to an expected value across individuals, ignoring cluster membership, if clusters sizes are equal. This implies that corresponding sample estimates of Symmetry (11) can only be a precise equality when all clusters are of equal sizes. To summarize, Symmetry (11) is a precise equality assuming cluster size is irrelevant to population parameters, but the extent to which sample estimates are precisely equal is likely dependent on the extent to which clusters are of equal size. Importantly, cluster sizes are not equivalent in typical empirical applications, therefore Symmetry (11) will not generally be an equivalence in practice.

Symmetry of factor loadings

Next, we consider Symmetries (11) and (13) for matrices of factor loadings. A deeper understanding of potential precursors to inequality can be gained by considering model-implied covariance structures. Given the relation between total effects and un-conflated level-1 and level-2 effects, the following is true:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{\theta}) &= \boldsymbol{\Sigma}_B(\boldsymbol{\theta}) + \boldsymbol{\Sigma}_W(\boldsymbol{\theta}) \\ &= \boldsymbol{\Lambda}_B\boldsymbol{\Psi}_B\Lambda_B' + \boldsymbol{\Lambda}_W\boldsymbol{\Psi}_W\boldsymbol{\Lambda}_W{}' + \boldsymbol{\Theta}_{\epsilon_{Bj}} + \boldsymbol{\Theta}_{\epsilon_{Wij}} \qquad (17)\end{aligned}$$

Further, using matrices defined in Equation (9), Equation (3) can be re-expressed as follows:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}_G\boldsymbol{\Psi}_G\boldsymbol{\Lambda}_G' + \boldsymbol{\Lambda}_S\boldsymbol{\Psi}_S\boldsymbol{\Lambda}_S' + \boldsymbol{\Theta}_\epsilon \qquad (18)$$

Where $\boldsymbol{\Psi}_G$ is a scalar depicting the variance of the general factor and $\boldsymbol{\Psi}_S$ is a diagonal matrix containing specific factor variances.

Equations (17) and (18) give rise to the following symmetries:

$$\boldsymbol{\Lambda}_B\boldsymbol{\Psi}_B\boldsymbol{\Lambda}_B' \approx \boldsymbol{\Lambda}_G\boldsymbol{\Psi}_G\boldsymbol{\Lambda}_G' \qquad (19)$$

$$\boldsymbol{\Lambda}_W\boldsymbol{\Psi}_W\boldsymbol{\Lambda}_W' \approx \boldsymbol{\Lambda}_S\boldsymbol{\Psi}_S\boldsymbol{\Lambda}_S' \qquad (20)$$

And equality:

$$\boldsymbol{\Theta}_{\epsilon_{Bj}} + \boldsymbol{\Theta}_{\epsilon_{Wij}} = \boldsymbol{\Theta}_\epsilon \qquad (21)$$

Symmetries (19) and (20) do not imply precise equality at the matrix or element level. In fact, we intentionally did not further decompose these by individual matrices, because equality of $\boldsymbol{\Psi}_S$ and $\boldsymbol{\Psi}_W$ is possible only in the specific and unlikely case of orthogonal level-1 factors. To clarify, bifactor models assume item covariances are completely captured by the general and specific factors, therefore specific factors do not covary. That is, $\boldsymbol{\Psi}_S$ is a diagonal matrix. In multilevel factor models, level-1 factors are typically allowed and expected to covary, meaning off-diagonal elements of $\boldsymbol{\Psi}_W$ likely will be non-zero. This implies that $\boldsymbol{\Psi}_S$ is equivalent to $\boldsymbol{\Psi}_W$ if and only if the off-diagonal elements of $\boldsymbol{\Psi}_W$ are zero, or level-1 factors are orthogonal, which is almost never observed in practice.

It is well known in structural equation modeling, that if parameters are erroneously fixed to zero, the unmodeled paths will bias parameter estimates in different parts of the model. Symmetry (20) suggests this phenomenon will likely bias elements of $\boldsymbol{\Lambda}_S$ inhibiting the possibility of precise equality. More technically, when off diagonal elements of $\boldsymbol{\Psi}_S$ are fixed to zero to identify the bifactor model, these unmodeled level-1 covariances will likely bias factor loading estimates. This could occur in one of three ways. Loadings for the specific factor could be inflated and loadings for the general factor could be attenuated, reducing overall support for the erroneous bifactor model in favor of a correlated factors solution where the "specific" factors are interpreted at level-1. Alternatively, loadings for the specific factor could be attenuated and loadings for the general factor could be inflated, increasing overall support for the erroneous bifactor model and resulting in a bifactor solution where the "specific" and "general" factors are interpreted at level-1. Finally, it is possible that there will be no discernible consequences of this misspecification. We aim to determine the behavior of this misspecification in the subsequent simulation demonstration.

Methods

Thus far, we have offered an analytically-guided discussion of the overlap between single-level bifactor models and multilevel measurement models fit to identically structured data. This has led to a series of symmetries, and motivated postulates regarding the extent to which these symmetries should not indicate precise equality in commonly encountered data structures. We explore these next.

Simulation design

We manipulated three key design factors: (1) level-1 factor correlations; (2) manifest variable ICCs; and (3) level-specific manifest variable communalities. Population

generating values were determined using a combination of findings from previously cited scale validation literature (Bae et al., 2020; Bae & DeBusk-Lane, 2019; Bartroli et al., 2022; Bianchi & Schonfeld, 2021; Calderon et al., 2018; Calvete et al., 2023; Caspi et al., 2024; Cuesta-Vargas et al., 2018; Dierendonck et al., 2020; Disabato et al., 2019; Maghsoodi et al., 2023; Ondé et al., 2022; Perera et al., 2018; Perreira et al., 2018; Quattrone et al., 2019, 2019; Savahl et al., 2023; Scherer et al., 2017; Shuck et al., 2019; Tiego et al., 2020; Torres-Vallejos et al., 2021; Verkuilen et al., 2021), additional results regarding ICCs commonly encountered in educational research (Zopluoglu, 2012), and the empirical data, Trends in International Mathematics and Science Study (TIMSS), utilized in the subsequent empirical example. All simulated data were generated using R (R Core Team, 2024) and factor models were fit using Mplus (Muthén & Muthén, 2017) in conjunction with the MplusAutomation R package (Hallquist & Wiley, 2018).

Data were generated from a multilevel factor model with one factor at level-2 and four factors at level-1, consistent with Figure 2. The level-2 population generating process was held constant across all simulation conditions. For simplicity, a single standardized latent variable was simulated with a total of 24 continuous indicators across 250 clusters, a similar number of clusters as was available for the subsequent real data analysis. Level-2 communalities were set to 0.64 throughout, implying unstandardized and standardized loadings of 0.8.

At level-1, the 24 observed indicators were evenly distributed across four standardized latent variables, inducing cross-level configural non-invariance. Here, two sets of level-1 inter-factor correlations were specified: 0.35 and 0.70. Non-zero off-diagonal elements of $\mathbf{\Psi}_W$ were chosen to explore the behavior of these unmodeled correlations in the misspecified bifactor model. We used a pseudo-random process to simulate unequal cluster sizes. Cluster sizes were drawn from a uniform distribution ranging from 2 to 74 (the minimum and maximum cluster size in the TIMSS dataset), such that the sum of observations led to a total level-1 sample size of 7,500. To improve efficiency of this process, cluster sizes were sampled in strata such that 60% were drawn from a uniform distribution ranging from 2 to 36 and 40% were drawn from a uniform distribution ranging from 37 to 74. The same set of unbalanced clusters were then continually used for all data simulated to consist of unequal cluster sizes.

We also varied indicator ICCs and level-1 communalities. Notably, these conditions were not originally considered when postulating criteria for symmetry inequalities; however, both are highly relevant to MCFA and bifactor models, respectively. First, indicator ICCs represent the extent to which cluster membership impacts observed variance and induces dependencies. As such, we considered three indicator ICCs represented in literature: 0.10, 0.30, and 0.50. In addition, bifactor structures may be preferred to non-hierarchical factor structures when standardized loadings on the general factor are relatively large. Because of this, although level-2 indicator communalities were held constant across all simulated datasets, level-1 indicator communalities were simulated with one of two selected communalities. In one condition, level-1 communalities were set to 0.64, or set equal to level-2 communalities. In the other condition, level-1 communalities were reduced to .36. We determined the size of raw loadings by solving for these within each ICC condition, holding communality constant at its predetermined value. A summary is presented in Table 1.

Table 1. Summary of simulation design.

ICC	Level-1 Communality	Unstandardized level-1 factor loading
0.10	0.36	1.80
	0.64	2.40
0.30	0.36	0.92
	0.64	1.22
0.50	0.36	0.60
	0.64	0.80

Taken together, this process resulted in a fully-crossed $2_{factor.correlation} \times 3_{ICC} \times 2_{level-1.communality}$ simulation design, or 12 design cells. Within each design cell, 500 datasets were simulated. Finally, within each of the dataset, two models were estimated: a misspecified single-level bifactor model and a properly specified MCFA.

Outcomes of interest

As stated, the primary goal of this simulation study was to evaluate the behavior of parameter estimates in the matrices involved in Symmetries (10), (11), and (13) when they are not expected to imply equality. We therefore extracted parameter estimates directly from each fitted measurement model, including all factor loadings and item intercepts from the unstandardized solutions and all standardized factor loadings.

A number of metrics were then computed using parameter estimates. Specifically, we computed raw differences of corresponding elements of matrices in Symmetries (10), (11), and (13). That is, for a given item $y_{k_{ij}}$ we computed the difference between intercepts from the MCFA and bifactor solution in

accordance with Symmetry (10), $\mu_k - \nu_k$, the difference between factor loadings at level-2 and general factor loadings in accordance with Symmetry (11) $\lambda_{k_B} - \lambda_{k_G}$, and the difference between factor loadings at level-1 and the specific factor loadings in accordance with Symmetry (13), $\lambda_{k_W} - \lambda_{k_S}$.

In addition, we used graphical analyses to evaluate the distribution of factor loadings from both models in relation to true population generating values. For example, for a given item $y_{k_{ij}}$ we plotted distributions of parameter estimates λ_{k_B} and λ_{k_G} against the true population generating factor loading at level-2.

Finally, we examined patterns of standardized loadings from the bifactor solutions to evaluate the extent to which an analyst might find empirical support for the misspecified bifactor model when structural clustering is not decomposed. For one, large standardized loadings on the general factor are often used as evidence in support of a general factor. In addition, large standardized loadings on the general factor in conjunction with diminished standardized loadings on specific factors will increase the size of many indices presented in Rodriguez et al. (2016) which are recommended to be used in place of standard fit indices to determine the necessity of a bifactor structure. In sum, if bifactor solutions result in relatively large standardized loadings on the general factor and relatively small standardized loadings on the specific, multilevel measurement models following the structure in Figure 2 will more successfully masquerade as bifactor models for all practical purposes.

Results

Intercept symmetry

As expected, Symmetry (10) was not a precise equality within a given data set due to differences in cluster size. However, the average intercept estimate from the bifactor solution within each design cell was equivalent to the population generating level-2 intercept. More technically, level-2 item intercepts were set to zero in all population generating structures, and within each simulation design cell, average intercept estimates produced from bifactor models were zero, within rounding to two decimal places. Therefore, even though the MCFA and bifactor solutions produced different intercept estimates due to variability in cluster size within a given sample, intercepts in bifactor solutions were able to correctly reproduce population generating values on average.

We computed correlation estimates between each $\hat{\mu}_k$ and $\hat{\nu}_k$ as a measure of overall deviations from equality. These are summarized in Figure 3.

Intercept estimates across the two models were most closely related when ICCs were small and became less related as ICCs increased. Specifically, when population generating ICCs were 0.10, correlations between $\hat{\mu}_k$ and $\hat{\nu}_k$ for a given $y_{k_{ij}}$ were as large as 0.93. This decreased to as low as 0.84 when ICCs were 0.50.

Factor loading symmetry

We now turn to Symmetries (11) and (13) regarding factor loadings. Similar to intercept estimates, we

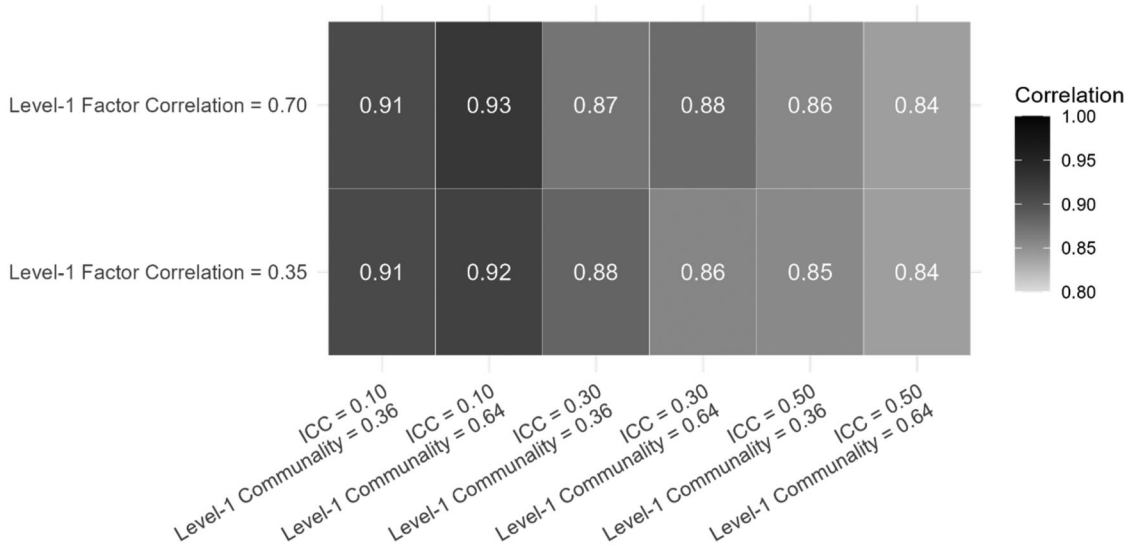

Figure 3. Heatmap of intercept correlations across simulation conditions.
Notes: correlation estimates are the Pearson product moment correlations between intercept estimates from the bifactor solution for each item $y_{k_{ij}}$ and associated level-2 intercept estimates from the MCFA.

Table 2. Evaluation of Symmetries (12) and (14) across level-1 factor correlation conditions.

	Level-1 Factor Correlations	
	0.35	0.70
	M (SD)	M (SD)
$\lambda_{k_{between}} - \lambda_{k_{General}}$	0.25 (0.13)	0.58 (0.29)
$\lambda_{k_{within}} - \lambda_{k_{Specific}}$.	−0.34 (0.27)	−0.58 (0.43)

computed the difference between associated elements of matrices in Symmetries (11) and (13). Specifically, for each item $y_{k_{ij}}$ we computed the difference between level-2 factor loadings and general factor loadings, $\hat{\lambda}_{k_B} - \hat{\lambda}_{k_G}$ and differences between level-1 factor loadings and specific factor loadings $\hat{\lambda}_{k_W} - \hat{\lambda}_{k_S}$.

Unlike intercept estimates, precise equalities when averaging across all sample estimates were not observed. Table 2 summarizes differences between these estimates within each data set, across the varying level-1 factor correlation conditions.

Next, we further explored patterns in the behavior of factor loadings across the two models. Figure 4 provides a detailed summary of distributions of factor loading within all simulation conditions.

The factor loadings in the bifactor model across Symmetries (11) and (13) did not lead to precise equalities, and the extent of this difference was dependent on the size of unmodeled level-1 correlation. Interestingly, a consistent pattern emerged. General factor loadings were higher than level-2 population generating factor loadings, and specific factor loadings were lower than level-1 population generating factor loadings.

Importantly, Figure 4 references differences in distributions of unstandardized factor loadings, which differed across simulation conditions in an effort to manipulate ICCs while holding communalities constant. Therefore, we turn to standardized factor loadings next to contextualize differences across conditions. Further, we explore insights into potential patterns in standardized factor loadings, as these are often used directly, or indirectly when using indices presented in Rodriguez et al. (2016), to make decisions regarding whether to retain a bifactor model over a correlated factors model.

Standardized factor loadings

Before presenting results regarding standardized factor loadings, we offer a brief analytical insight to contextualize findings. Importantly, it is expected that all standardized loadings from the bifactor solution will be attenuated compared to standardized loadings from the MCFA. This is because the MCFA decomposes variance into respective within- and between-components. Therefore, level-2 factor loadings are standardized with respect to the proportion of item variance due to level-2, whereas level-1 factor loadings are standardized with respect to the proportion of item variance due to level-1. This decreases the size of the denominator in the process of standardization, thereby increasing the absolute value of the standardized estimate, when comparing solutions to a model that conflates level-1 and level-2 effects. As a result, Figure 5 shows that standardized loadings from the bifactor solution are consistently lower than those from the MCFA. This is a natural byproduct of the difference inherent in standardization formulae underlying these modeling procedures. With this in mind, we focus our discussion of standardized loadings by comparing loadings within a modeling framework as opposed to across modeling frameworks (e.g., comparing general factor loadings to specific factor loadings within the bifactor solution).

Figure 5 elucidates an interesting pattern when comparing the size of the standardized factor loadings on the general factor compared to specific factors in the bifactor solution. To explain, we will focus on examining patterns observed in the right-side panel, wherein communalities at level-1 and level-2 were equivalent so that direct comparisons are interpretable. Specifically, when ICCs were set to 0.10, standardized loadings on the general factor were, on average, smaller than standardized loadings on the specific factors when level-1 factor correlations were 0.35 ($M_{\lambda_G} = 0.52$, $SD_{\lambda_G} = 0.01; M_{\lambda_S} = 0.61, SD_{\lambda_S} = 0.01$). This pattern shifted for larger level-1 factor correlations, wherein loadings on the general factor were larger ($M_{\lambda_G} = 0.68, SD_{\lambda_G} = 0.01; M_{\lambda_S} = 0.42, SD_{\lambda_S} = 0.01$). When ICCs were 0.30, standardized general factor loadings were larger with factor correlations of 0.35 ($M_{\lambda_G} = 0.59, SD_{\lambda_G} = 0.02; M_{\lambda_S} = 0.54, SD_{\lambda_S} = 0.01$) and 0.70 ($M_{\lambda_G} = 0.71$, $SD_{\lambda_G} = 0.01$; $M_{\lambda_S} = 0.37$, $SD_{\lambda_S} = 0.02$). This pattern continued with large ICCs of 0.50, wherein general factor loadings were consistently larger than specific factor loadings, regardless of level-1 factor correlations.

When level-1 communalities were set to be lower than level-2 item communalities, similar patterns were observed, but the differences between standardized loadings on the general factor and on specific factors were more pronounced, in favor of larger loadings on the general factor.

To summarize, standardized loadings across Symmetries (11) and (13) were consistently nonequivalent when single-level bifactor models were employed as opposed to the full MCFA, as expected. The extent of

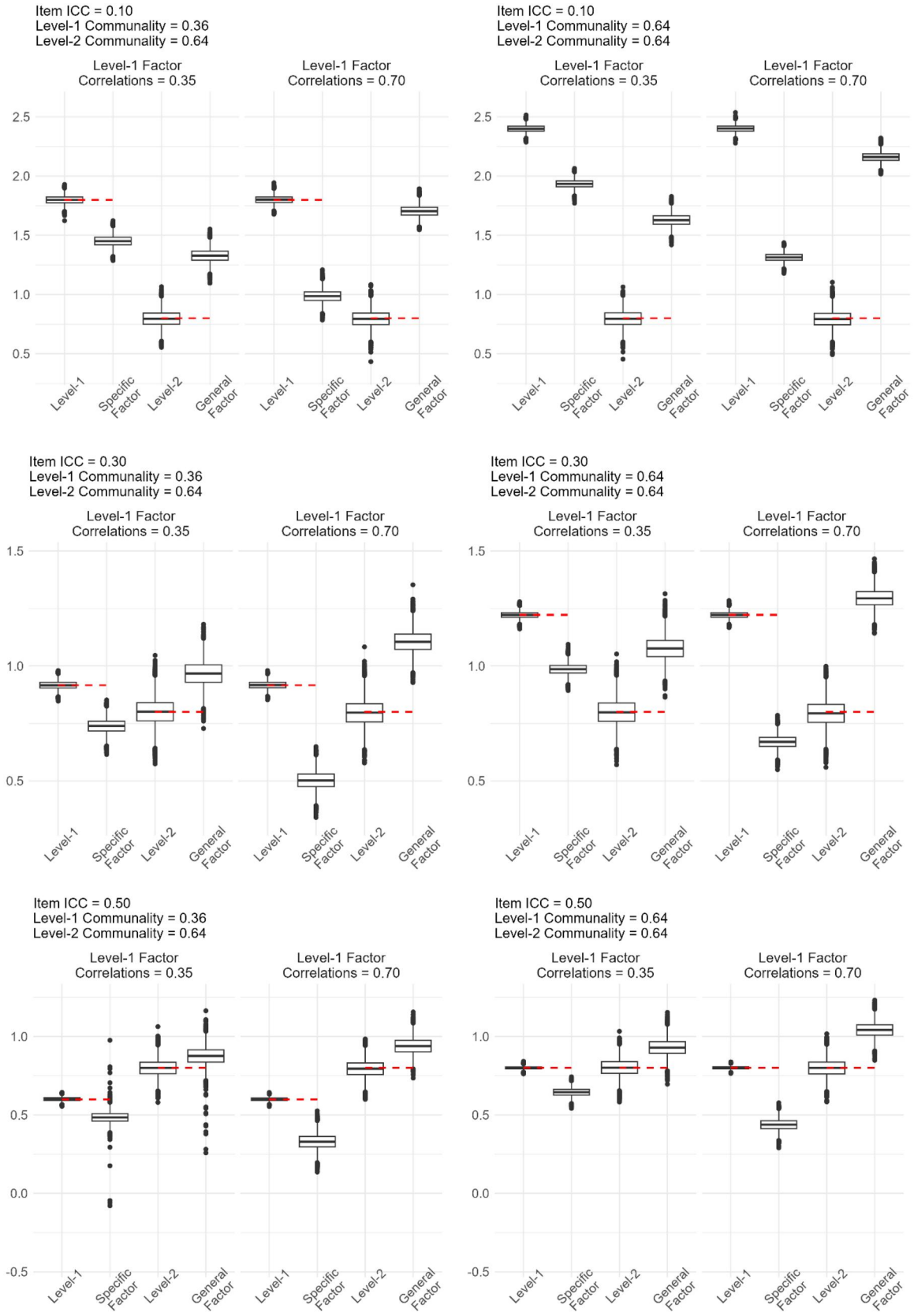

Figure 4. Summary of unstandardized factor loadings. *Notes*. Dashed red lines indicate population generating values.

this difference depended on a combination of both level-1 factor correlations and ICCs. With low level-1 factor correlations and low ICCs, general factor loadings were generally smaller than specific factor loadings. With high level-1 factor correlations and/or high ICCs, general factor loadings were generally larger than

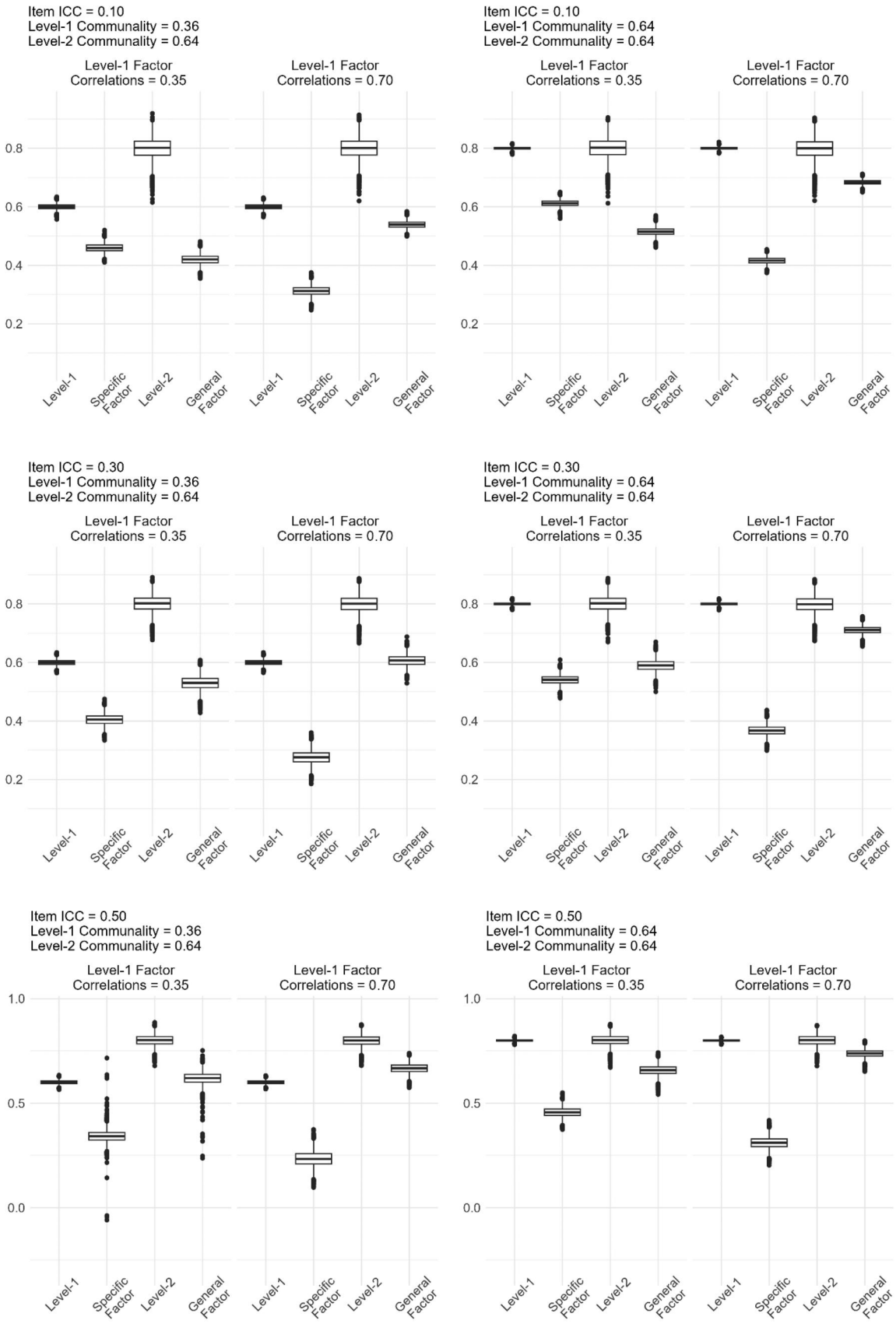

Figure 5. Summary of standardized factor loadings.

specific factor loadings. Taken together, in the majority of simulation conditions the unmodeled level-1 correlations in the misspecified bifactor model led to inflated standardized loadings on the general factor, increasing the likelihood an applied researcher would retain the bifactor solution if level-1 and level-2 effects are not decomposed.

Empirical demonstration

Next, we present findings from empirical data to demonstrate multilevel factor models masquerading as bifactor models in practice. The purpose of this demonstration is to show how researchers employing measurement models that conflate level-1 and level-2 processes due to clustering of individuals in larger social/organizational units may find empirical support for a bifactor solution, when the data actually conform to a multilevel factor model with fewer level-2 factors than level-1 factors. In addition, we will apply insights gained from the previous simulation study to explain results.

Sample and measures

Data come from the Trends in International Mathematics and Science Study (TIMSS; U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2021). We analyzed a subset of all available data from 2023 data, focusing on eighth grade students in the United States. The data consisted of a total of 8,074 unique students clustered in 211 schools.

Further, we selected the TIMSS 2023 Achievement Scales for analysis. These are separated into two achievement domains, mathematics and science. For the eighth-grade datasets, mathematics achievement is further decomposed into four content domains that are assigned a single numeric score: numbers, algebra, geometry, and data and probability. Science also includes four content domain scores for biology, chemistry, physics, and earth science.

This dataset was particularly relevant given theoretically supported hypotheses regarding construct overlap. Specifically, while math and science achievement represent separate academic domains, it is likely these are not distinct constructs at the school level, and instead represent a school-level STEM (science, technology, engineering, and mathematics) education factor. The decomposition into separate math and science achievement factors may be of relevance only when considering the student-level measurement structure.

In practice, however, measurement models that conflate potentially non-invariant level-1 and level-2 measurement models are not often utilized. For example, an analyst may choose to address the clustering of students in schools through cluster-corrected standard errors; alternatively, an analyst may not be aware of the consequences of ignoring clustering altogether and may choose to not address the clustering of students in schools. Regardless of their reasoning, our simulation demonstration suggests this analyst is likely to find empirical support for a bifactor model in the form of large standardized loadings on a general factor that is a masquerade of a true level-2 factor due to clustering of students within schools.

To demonstrate this masquerade, our analytic strategy involved fitting both an MCFA and a bifactor CFA to the data. We begin with a summary of findings from the multilevel analysis followed by a summary using non-multilevel CFA. All models were fit using Mplus (Muthén & Muthén, 2017) and model identification was achieved by standardizing the scales of latent variables.

MCFA

Achievement score ICCs for these data ranged from 0.27 to 0.36, suggesting a meaningful proportion of variance due to school membership, and further suggesting the need to decompose school- and student-level effects. We began by fitting two competing models, one with separate science and mathematics achievement factors at both the student- and school-level, and one with separate factors at the student-level and a single STEM education factor at the school-level. While there were not substantial differences in model fit between the two models, when separate school-level factors were estimated, the correlation between these factors was 0.99. Given hypothesized configural cross-level non-invariance with the existence of a STEM education factor at the school level, we retained the model with a single level-2 STEM education factor ($\mathrm{CFI} = 0.99; \mathrm{TLI} = 0.99;\ \mathrm{RMSEA} = 0.05; \mathrm{SRMR}_{between} = 0.01, \mathrm{SRMR}_{within} = 0.02$).

Bifactor CFA

Next, we analyzed the same data using a factor model that does not decompose student- and school-level effects. Data for both the MCFA and bifactor CFA were in "wide" form, that is, no transformations were required to move from one modeling framework to the other given the source of clustering was students nested in schools. To account for the impact of clustered data

Table 3. Comparison of MCFA and bifactor solutions in empirical data.

	MCFA Estimate (Standardized Estimate)		Bifactor Solution Estimate (Standardized Estimate)	
	Level-1		Specific Factors	
Math				
Numbers	0.75 (0.97)		0.47 (0.48)	
Algebra	0.76 (0.95)		0.46 (0.45)	
Geometry	0.68 (0.94)		0.38 (0.41)	
Data and Probability	0.75 (0.94)		0.41 (0.41)	
Science				
Biology	0.81 (0.96)		0.27 (0.27)	
Chemistry	0.74 (0.93)		0.09 (0.10)	
Physics	0.83 (0.95)		0.17 (0.16)	
Earth Science	0.82 (0.94)		0.21 (0.20)	
Science with Math	0.81 (0.81)		–	
	Level-2		General Factor	
	Factor Loading	Item Intercept	Factor Loading	Item Intercept
STEM Education				
Numbers	0.57 (0.99)	4.86	0.83 (0.86)	4.88
Algebra	0.60 (0.99)	4.91	0.86 (0.86)	4.93
Geometry	0.53 (0.99)	4.81	0.79 (0.87)	4.82
Data and Probability	0.58 (0.99)	4.92	0.86 (0.87)	4.92
Biology	0.51 (0.99)	5.16	0.94 (0.94)	5.18
Chemistry	0.49 (0.99)	5.05	0.90 (0.95)	5.06
Physics	0.54 (0.99)	5.14	0.98 (0.95)	5.16
Earth Science	0.54 (0.99)	5.09	0.97 (0.93)	5.11

on standard errors, we used robust maximum likelihood with corrected standard errors. We began with a correlated factors model including unique mathematics achievement and science achievement factors, but without a general factor, to determine whether a bifactor structure would be empirically justified in comparison to a simpler model.

In this model, the science and math factors correlated 0.87. The solution fit well ($\mathrm{CFI} = 0.99$; $\mathrm{TLI} = 0.99$; $\mathrm{RMSEA} = 0.08$; $\mathrm{SRMR} = 0.01$) but improved when moving to the bifactor model ($\mathrm{CFI} = 1.00$; $\mathrm{TLI} = 0.99$; $\mathrm{RMSEA} = 0.06$; $\mathrm{SRMR} < 0.01$). Given improvements in fit are expected when comparing a correlated factors solution to a bifactor model and may not necessarily imply the bifactor solution is optimal (Murray & Johnson, 2013; Reise, 2012; Reise et al., 2023; Rodriguez et al., 2016) we further computed indices suggested by Rodriquez et al. (2016) to provide a more holistic exploration of the scale's psychometric properties. Coefficient omega, coefficient omega hierarchical, and explained common variance for the general factor were large ($\omega = 0.99; \omega_H = 0.93; ECV = 0.88$). Most item ECVs (I-ECVs) exceeded 0.80, a commonly used cutoff to support unidimensionality (Stucky et al., 2013). Taken together, empirical evidence supported the bifactor solution when clustering was addressed with cluster-corrected standard errors but level-1 and level-2 effects from clustering of students in schools were not decomposed.

Parameter estimates across Symmetries (10), (11), and (13) generally followed predictable patterns utilizing findings from the previous simulation study. Contextually, these data most resembled the simulation condition with mid-level ICCs (0.35), high level-1 factor correlations (0.7), and smaller level-1 item communalities compared to level-2 item communalities. As such, based on findings from the simulation study, we expected item intercepts to be roughly similar across the models but not equivalent. This was found to be true. Small differences were observed across intercept estimates. Further, we expected the bifactor solution to produce standardized specific factor loadings lower than what is obtained in the MCFA and standardized general factor loadings higher than what is obtained in the MCFA. This was also found to be true. Importantly, based on findings from our simulation demonstration, larger loadings on the general factor compared to specific factors in the bifactor solution could be explained as the result of misspecified inter-factor correlations. In this empirical example, these likely produced larger values of ω, ω_H, and ECV, increasing support for the bifactor solution when level-1 and level-2 effects were conflated. Finally, unstandardized loadings on the specific factors were generally small, with one approaching zero ($\lambda_S = 0.09$), suggesting poor local fit. All unstandardized loadings in the MCFA were substantial, consistent with theoretical expectations. Findings are summarized in Table 3.

Conclusion

The purpose of this investigation was to demonstrate that a particular modeling approach that does not decompose level-1 and level-2 effects due to clustering of individuals in social and/or organizational units can result in erroneous conclusions regarding underlying measurement structures when fit to clustered data. Specifically, we considered multilevel measurement structures with fewer level-2 factors compared to level-1 factors, a measurement structure common in multilevel applications (Kim et al., 2016), and explored how these may be misrepresented as bifactor solutions, another common structure in practice (Eid et al., 2017; Reise, 2012; Reise et al., 2016; Rodriguez et al., 2016), when level-1 and level-2 effects are not decomposed. This investigation was motivated by noting an evident structural symmetry in the manifest variable equations and model-implied covariance formulae underlying both the MCFA with a single level-2 factor and bifactor models.

These symmetries led to a series of postulates that were further tested using simulation methodology. According to our symmetries, precise equivalencies across the modeling frameworks, without restructuring of data, are contingent on cluster sizes being equivalent and level-1 factors being orthogonal, both of which are unlikely to occur in practice. Therefore, the misspecified bifactor solution should lead to different parameter estimates than the properly specified multilevel measurement model, but the nature of these differences required further analysis. Therefore, we conducted a simulation study and used an empirical example to evaluate sample behavior.

One of the most practically important findings observed in the simulation study and in the subsequent empirical data demonstration concerned the patterns of differences among standardized loadings in the bifactor solution. Importantly, in most simulation conditions and in the empirical data demonstration, the bifactor solution produced larger general factor loadings than specific factor loadings. This is critical because larger general factor loadings compared to specific factor loadings may be used as empirical evidence in support of the existence of a bifactor solution over a correlated factors solution.

Relatedly, the empirical demonstration showed that standard fit indices are not always effective at determining if the population-generating process is a multilevel factor model without a bifactor structure or a classic single-level bifactor model. The bifactor model fit the data well. Metrics specifically recommended for use to offset the proclivity of bifactor models to fit well further supported the bifactor solution. Further, in both the simulation and empirical demonstration most unstandardized and standardized loadings in the classic bifactor model were substantial. There was the exception of a single unstandardized loading on a specific factor in the empirical example approaching zero. Vanishing specific factor loadings have been noted to occur when bifactor models are erroneously applied to data where specific factors are not interchangeable—a characteristic of the TIMSS example (Eid et al., 2017). That said, vanishing specific factor loadings were not consistently found on average in the simulation demonstration. This suggests that typical anomalous results that occur due to inappropriate application of classic bifactor solutions are not a consistent artifact of the masquerade presented here. This is expected in a model masquerade, but concerning in that these findings undermine the interpretability of local and global fit.

Regarding the observed patterns of standardized loadings, we believe this is generally explainable by returning to the model-implied covariance structures. Specifically, bifactor models require a diagonal covariance matrix for latent factors. That is, the specific factors are not allowed to covary. If the true data-generating process involves factor covariances at level-1, these covariances are likely to bias factor loadings matrices to maximize the likelihood of the observed item covariance matrix. This explains why the degree of differences in factor loadings from the bifactor solution across Symmetries (11) and (13) differed substantially from factor loadings from the MCFA with larger level-1 factor correlations.

In sum, findings suggest that, practically speaking, multilevel measurement structures with fewer level-2 factors than level-1 factors can masquerade as bifactor models when clustering of individuals in social/organization units is not decomposed *via* a multilevel factor model. It is important to explain why this is consequential in practice. Specifically, if a CFA can correctly determine that the underlying latent variable structure involves a single factor relating all items and a series of specific factors, why would it be necessary to consider the MCFA? The answer to this inquiry lies in model interpretation. General factors and specific factors are substantively different from level-2 and level-1 factors. Namely, general factors and specific factors can both be applied to individuals whereas level-1 factors are unique to individuals and level-2 factors are unique to clusters. As such, if a general factor emerges in measurement modeling, and this general factor emerges solely due to a single factor that exists at the cluster-level, interpretation of the general factor at the individual-level is theoretically invalid. Valid interpretations of the factors can be made only when they are accurately applied to the

correct level of the hierarchy, and the correct level of the hierarchy can be determined only in the MCFA.

This leads to the primary solution to this masquerade: we encourage researchers with clustered data and nontrivial manifest variable ICCs to consider MCFA as an alternative modeling approach to the bifactor CFA. An alternative explanation for empirical support of a bifactor structure is indeed a multilevel data-generating model with a single level-2 factor, multiple correlated level-1 factors, and no bifactor structure. In some scenarios this multilevel structure might better describe the true data-generating process and might better map onto theory, and this can thoroughly be empirically investigated only using multilevel measurement models.

This solution, however, applies only when the source of clustering is known and measured. There are likely data structures in which clustering in a sample or population exists but is unknown and/or unmeasured. It is therefore theoretically tenable that any bifactor structure might be more accurately described as a multilevel structure with a single factor at level-2 and distinct factors at level-1, regardless of known or measured sources of clustering. Because of potential consequences of unmodeled clustering, we encourage all researchers finding ubiquitous support for bifactor solutions in clustered data to consider whether or not applying the general factor at the individual-level is theoretically supported. If not, the general factor may be a byproduct of different factor structures at different levels of analysis, or cross-level non-invariance, had those levels of analysis been unconflated.

This leads to a discussion of limitations and suggestions for future research. In accordance with the establishment of structural symmetries, we did not further consider multilevel population generating structures involving a bifactor structure at level-1 and a single factor at level-2. This creates an additional series of loadings complicating established symmetries by removing equivalence of matrix dimensions; however, it may be the case that these structures are even more prone to a potential masquerade, though one in which a version of the "general" factor is interpretable at level-1. Flexibility of measurement models at level-1 and level-2 is a benefit of the MCFA and future research should continue to consider alternative potential specifications that may produce erroneous structures when level-1 and level-2 effects are not decomposed.

In addition, future research should place a greater emphasis on varied model evaluative metrics used in bifactor models, including, but not limited to, indices suggested in Rodriguez et al. (2016). Our study offered an analytically guided initial investigation into certain multilevel factor structures masquerading as bifactor models, demonstrated that this masquerade occurs, and established patterns of differences across parameter estimates. Future studies should use a more holistic series of outcomes, above and beyond standardized loadings, to determine the relative frequency with which a bifactor structure would be retained over other competing models that conflate level-1 and level-2 effects.

Finally, our study was limited by considering only bifactor models as a possible masquerade to multilevel measurement models, given their frequency in application and noted proclivity for misuse (Eid et al., 2017; Murray & Johnson, 2013; Reise, 2012; Reise et al., 2023; Rodriguez et al., 2016). There may be additional structures vulnerable to this masquerade. For one, bifactor models can be shown to be equivalent to hierarchical CFA models (Markon, 2019; Yung et al., 1999). Since hierarchical CFA allows direct paths between higher order factors and lower order factors but does not allow manifest variables to directly load on higher order factors, parameter estimates are likely to follow different patterns than what was observed in Figures (4) and (5). Further, we did not consider a comparison between multilevel measurement models and the bifactor(S-1) or bifactor(S*I−1) models, which are more appropriate when the sampling procedure of specific factors and general factor is not a two-level sampling procedure (Eid et al., 2017). These models allow some, but not all, specific factor correlations to be estimated; therefore, if they are erroneously applied to data whose population structure mirrors what was considered here, the patterns of differences across Symmetries (11) and (13) are likely to be different. We also limited the investigation to multilevel models with fewer factors at level-2 than at level-1 given the parallel with bifactor models, but results may be similar in multilevel models with different structures. That is, a multilevel measurement model with an equivalent number of factors at level-1 and level-2 could still masquerade as a bifactor-like model with the same number of general factors as specific factors and the latter would still be invalid if factors are not interpreted at the correct level of analysis. Future research should consider alternative but similar factor structures, such as hierarchical CFA, the bifactor(S-1) or bifactor(S*I−1), or more complicated bifactor-like models to better understand the nature of additional potential masquerade.

In sum, we have offered evidence that certain bifactor solutions may, in fact, be manifestations of multilevel processes wherein fewer factors exist at the cluster-level compared to the individual-level (i.e., cross-level configural non-invariance). This initial investigation

offers insights to inform future methodological work considering the breadth of potential consequences of conflating level-1 and level-2 effects in the measurement and scale validation literature. Future research should continue to explore these overlaps, ensuring that measurement models accurately reflect the underlying data structures and inform valid theoretical construct interpretation, particularly in complex data structures.

References

Asparouhov, T., & Muthén, B. (2006). *Constructing covariates in multilevel regression*.

Bae, C. L., & DeBusk-Lane, M. (2019). Middle school engagement profiles: Implications for motivation and achievement in science. *Learning and Individual Differences*, *74*, 101753. https://doi.org/10.1016/j.lindif.2019.101753

Bae, C. L., DeBusk-Lane, M. L., & Lester, A. M. (2020). Engagement profiles of elementary students in urban schools. *Contemporary Educational Psychology*, *62*, 101880. https://doi.org/10.1016/j.cedpsych.2020.101880

Bartroli, M., Angulo-Brunet, A., Bosque-Prous, M., Clotas, C., & Espelt, A. (2022). The Emotional Competence Assessment Questionnaire (ECAQ) for children aged from 3 to 5 years: Validity and reliability evidence. *Education Sciences*, *12*(7), 489. https://doi.org/10.3390/educsci12070489

Bauer, D. J., & Cai, L. (2009). Consequences of unmodeled nonlinear effects in multilevel models. *Journal of Educational and Behavioral Statistics*, *34*(1), 97–114. https://doi.org/10.3102/1076998607310504

Belzak, W. C. M., & Bauer, D. J. (2019). Interaction effects may actually be nonlinear effects in disguise: A review of the problem and potential solutions. *Addictive Behaviors*, *94*, 99–108. https://doi.org/10.1016/j.addbeh.2018.09.018

Bianchi, R., & Schonfeld, I. S. (2021). Occupational depression, cognitive performance, and task appreciation: A study based on Raven's Advanced Progressive Matrices. *Frontiers in Psychology*, *12*, 695539. https://doi.org/10.3389/fpsyg.2021.695539

Brown, T. A. (2006). *Confirmatory factor analysis for applied research* (p. 475). The Guilford Press.

Cai, L., & Houts, C. R. (2021). Longitudinal analysis of patient-reported outcomes in clinical trials: Applications of multilevel and multidimensional item response theory. *Psychometrika*, *86*(3), 754–777. https://doi.org/10.1007/s11336-021-09777-y

Calderon, C., Jiménez-Fonseca, P., Ferrando, P. J., Jara, C., Lorenzo-Seva, U., Beato, C., García-García, T., Castelo, B., Ramchandani, A., Muñoz, M. M., Martínez De Castro, E., Ghanem, I., Mangas, M., & Carmona-Bayonas, A. (2018). Psychometric properties of the Shared Decision-Making Questionnaire (SDM-Q-9) in oncology practice. *International Journal of Clinical and Health Psychology: IJCHP*, *18*(2), 143–151. https://doi.org/10.1016/j.ijchp.2017.12.001

Calvete, E., Jiménez-Granado, A., & Orue, I. (2023). The revised Child-to-Parent Aggressions Questionnaire: An examination during the covid-19 pandemic. *Journal of Family Violence*, *38*(8), 1–14. https://doi.org/10.1007/s10896-022-00465-8

Caspi, A., Houts, R. M., Fisher, H. L., Danese, A., & Moffitt, T. E. (2024). The general factor of psychopathology (p): Choosing among competing models and interpreting p. *Clinical Psychological Science: A Journal of the Association for Psychological Science*, *12*(1), 53–82. https://doi.org/10.1177/21677026221147872

Castro-Alvarez, S., Tendeiro, J. N., Meijer, R. R., & Bringmann, L. F. (2022). Using structural equation modeling to study traits and states in intensive longitudinal data. *Psychological Methods*, *27*(1), 17–43. https://doi.org/10.1037/met0000393

Cuesta-Vargas, A. I., Neblett, R., Chiarotto, A., Kregel, J., Nijs, J., Van Wilgen, C. P., Pitance, L., Knezevic, A., Gatchel, R. J., Mayer, T. G., Viti, C., Roldan-Jiménez, C., Testa, M., Caumo, W., Jeremic-Knezevic, M., & Luciano, J. V. (2018). Dimensionality and reliability of the central sensitization inventory in a pooled multicountry sample. *The Journal of Pain*, *19*(3), 317–329. https://doi.org/10.1016/j.jpain.2017.11.006

Dierendonck, C., Milmeister, P., Kerger, S., & Poncelet, D. (2020). Examining the measure of student engagement in the classroom using the bifactor model: Increased validity when predicting misconduct at school. *International Journal of Behavioral Development*, *44*(3), 279–286. https://doi.org/10.1177/0165025419876360

Disabato, D. J., Goodman, F. R., & Kashdan, T. B. (2019). Is grit relevant to well-being and strengths? Evidence across the globe for separating perseverance of effort and consistency of interests. *Journal of Personality*, *87*(2), 194–211. https://doi.org/10.1111/jopy.12382

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods*, *22*(3), 541–562. https://doi.org/10.1037/met0000083

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989x.12.2.121.supp

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley.

Hox, J. J. (1993). Factor analysis of multilevel data: Gauging the Muthén model. In J. H. L. Oud & R. A. W. van Blokland-Vogelsang (Eds.), *Advances in longitudinal and multivariate analysis in the behavioral sciences* (Vol. 10, pp. 141–156). ITS.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 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*, 1640. https://doi.org/10.3389/fpsyg.2017.01640

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(2), 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: A Multidisciplinary Journal*, *21*(1), 31–39. https://doi.org/10.1080/10705511.2014.856694

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93. https://doi.org/10.1111/j.1745-3984.2001.tb01117.x

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

Kreft, I. G. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*(3), 385–393. https://doi.org/10.1037/0033-2909.107.3.385

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

Maghsoodi, A. H., Ruedas-Gracia, N., & Jiang, G. (2023). Measuring college belongingness: Structure and measurement of the Sense of Social Fit Scale. *Journal of Counseling Psychology*, *70*(4), 424–435. https://doi.org/10.1037/cou0000668

Mancl, L. A., Leroux, B. G., & DeRouen, T. A. (2000). Between-subject and within-subject statistical information in dental research. *Journal of Dental Research*, *79*(10), 1778–1781. https://doi.org/10.1177/00220345000790100801

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*(1), 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140. https://doi.org/10.1037/met0000078

Miller, M. D., & Burstein, L. (1981). *Multilevel properties of test items: An exploratory study*. CSE Report No. 181, Center for the Study of Evaluation, University of California, Los Angeles.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, *41*(5), 407–422. https://doi.org/10.1016/j.intell.2013.06.004

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, *54*(2), 638–645. https://doi.org/10.2307/3109770

Neuhaus, J. M., & McCulloch, C. E. (2006). Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(5), 859–872. https://doi.org/10.1111/j.1467-9868.2006.00570.x

Ondé, D., Jiménez, V., Alvarado, J. M., & Gràcia, M. (2022). Analysis of the structural validity of the reduced version of metacognitive awareness of reading strategies inventory. *Frontiers in Psychology*, *13*, 894327. https://doi.org/10.3389/fpsyg.2022.894327

Perera, H. N., Vosicka, L., Granziera, H., & McIlveen, P. (2018). Towards an integrative perspective on the structure of teacher work engagement. *Journal of Vocational Behavior*, *108*, 28–41. https://doi.org/10.1016/j.jvb.2018.05.006

Perreira, T. A., Morin, A. J. S., Hebert, M., Gillet, N., Houle, S. A., & Berta, W. (2018). The short form of the Workplace Affective Commitment Multidimensional Questionnaire (WACMQ-S): A bifactor-ESEM approach among healthcare professionals. *Journal of Vocational Behavior*, *106*, 62–83. https://doi.org/10.1016/j.jvb.2017.12.004

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*(2), 189–205. https://doi.org/10.1037/met0000052

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

Quattrone, D., Di Forti, M., Gayer-Anderson, C., Ferraro, L., Jongsma, H. E., Tripoli, G., La Cascia, C., La Barbera, D., Tarricone, I., Berardi, D., Szöke, A., Arango, C., Lasalvia, A., Tortelli, A., Llorca, P.-M., de Haan, L., Velthorst, E., Bobes, J., Bernardo, M., … Reininghaus, U, EU-GEI WP2 Group (2019). Transdiagnostic dimensions of psychopathology at first episode psychosis: Findings from the multinational EU-GEI study. *Psychological Medicine*, *49*(8), 1378–1391. https://doi.org/10.1017/S0033291718002131

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear modeling: Applications and data analysis methods* (2nd ed.). SAGE Publications Ltd.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*(6), 818–838. https://doi.org/10.1080/00273171.2016.1243461

Reise, S. P., Mansolf, M., & Haviland, M. G. (2023). Bifactor measurement models. In R. H. Hoyle (Ed.), *Handbook of structural equation modelling* (pp. 329–348). Guilford Press.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *The British Journal of Mathematical and Statistical Psychology*, *67*(1), 172–194. https://doi.org/10.1111/bmsp.12014

Savahl, S., Casas, F., & Adams, S. (2023). Considering a bifactor model of children's subjective well-being using a multinational sample. *Child Indicators Research*, *16*(6), 2253–2278. https://doi.org/10.1007/s12187-023-10058-6

Scherer, R., Tondeur, J., & Siddiq, F. (2017). On the quest for validity: Testing the factor structure and measurement invariance of the technology-dimensions in the Technological, Pedagogical, and Content Knowledge (TPACK) model. *Computers & Education*, *112*, 1–17. https://doi.org/10.1016/j.compedu.2017.04.012

Shuck, B., Alagaraja, M., Immekus, J., Cumberland, D., & Honeycutt-Elliott, M. (2019). Does compassion matter in leadership? A two-stage sequential equal status mixed method exploratory study of compassionate leader behavior and connections to performance in human resource development. *Human Resource Development Quarterly*, *30*(4), 537–564. https://doi.org/10.1002/hrdq.21369

Stapleton, L. M., McNeish, D. M., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist*, *51*(3–4), 317–330. https://doi.org/10.1080/00461520.2016.1207178

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

Stucky, B. D., Thissen, D., & Orlando Edelen, M. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, *37*(1), 41–57. https://doi.org/10.1177/0146621612462759

Tiego, J., Bellgrove, M. A., Whittle, S., Pantelis, C., & Testa, R. (2020). Common mechanisms of executive attention underlie executive function and effortful control in children. *Developmental Science*, *23*(3), e12918. https://doi.org/10.1111/desc.12918

Torres-Vallejos, J., Juarros-Basterretxea, J., Oyanedel, J. C., & Sato, M. (2021). A bifactor model of subjective well-being at personal, community, and country levels: a case with three Latin-American Countries. *Frontiers in Psychology*, *12*, 641641. https://doi.org/10.3389/fpsyg.2021.641641

Ulitzsch, E., Holtmann, J., Schultze, M., & Eid, M. (2017). Comparing multilevel and classical confirmatory factor analysis parameterizations of multirater data: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 80–103. https://doi.org/10.1080/10705511.2016.1251846

Verkuilen, J., Bianchi, R., Schonfeld, I. S., & Laurent, E. (2021). Burnout–depression overlap: Exploratory structural equation modeling bifactor analysis and network analysis. *Assessment*, *28*(6), 1583–1600. https://doi.org/10.1177/1073191120911095

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113–128. https://doi.org/10.1007/BF02294531

Zopluoglu, C. (2012). A cross-national comparison of intraclass correlation coefficient in educational achievement outcomes. *Journal of Measurement and Evaluation in Education and Psychology*, *3*(1), 242–278.

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*(2), 127–140. https://doi.org/10.1037/1089-2699.12.2.127