The Sampling Ratio in Multilevel Structural Equation Models: Considerations to Inform Study Design

Joseph M. Kush[1][*], Timothy R. Konold[1], and Catherine P. Bradshaw[1]

[1]University of Virginia, School of Education and Human Development

[*]Correspondence concerning this article should be addressed to Joseph Kush, Johns Hopkins Bloomberg School of Public Health, 624 N Broadway Room 841, Baltimore, MD 21205. Email: jkush1@jhu.edu

Abstract

Multilevel structural equation (MSEM) models allow researchers to model latent factor structures at multiple levels simultaneously by decomposing within- and between-group variation. Yet the extent to which the sampling ratio (i.e., proportion of cases sampled from each group) influences the results of MSEM models remains unknown. This paper explores how variation in the sampling ratio in MSEM impacts the measurement of Level-2 (L2) latent constructs. Specifically, we investigated whether the sampling ratio is related to bias and variability in aggregated L2 construct measurement and estimation in the context of doubly latent MSEM models utilizing a two-step Monte Carlo simulation study. Findings suggest that while lower sampling ratios were related to increased bias, standard errors, and RMSE, the overall size of these errors was negligible, making the doubly latent model an appealing choice for researchers. An applied example using empirical survey data is further provided to illustrate the application and interpretation of the model. We conclude by considering the implications of various sampling ratios on the design of MSEM studies, with a particular focus on educational research.

Clustered data structures can present challenges for general linear models that assume independence of observations (Bentler & Chou, 1987; Raudenbush & Bryk, 2002). Yet clustered data structures can provide an opportunity for understanding how variables are related at more than one level. Multilevel modeling (MLM) adjusts for the violation of the homoskedasticity (or exogeneity) assumption of single-level linear regression, in which residual variance are no longer independent (Bentler & Chou, 1987; Raudenbush & Bryk, 2002; Rabe-Hesketh & Skrondal, 2008). MLM is frequently used to account for clustered data, where data are often clustered within higher level units, and thus continues to grow in popularity among applied researchers in the social and behavioral sciences (Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). For instance, students may be clustered within schools, teachers may be clustered within districts, and individuals may be clustered within neighborhoods. By partitioning variation at different levels, multilevel modeling allows researchers to explore relationships among variables at two or more levels, and at the same time control for violations of independence that can have an adverse impact on the standard errors of model parameter estimates.

While observed variable multilevel analysis has a long-standing history in the social sciences, multilevel structural equation modeling (MSEM) continues to make substantial progress both in theory and application (see Hox & Maas, 2001; Jia & Konold, 2019; Muthén & Asparouhov, 2009; Preacher et al., 2010). MSEM is useful in accounting for measurement and/or sampling error, where Level-1 (L1) and Level-2 (L2) substantive latent constructs are estimated from a set of manifest observed variables. MSEM has spurred a number of theoretical and methodological research investigations that have focused on the performance of these models in applied settings. Some examples include estimating reliability of the L2 construct (Bliese, 2000),

as well as formulating latent substantive constructs on the basis of latent-manifest variables (Lüdtke et al., 2011; Marsh et al., 2009).

One critical issue that has been underexamined in the MSEM literature is the sampling ratio. The sampling ratio represents the proportion of L1 cases sampled out of the total number of L1 cases within each L2 group that are potentially available. Often in educational and social sciences, researchers utilize cluster sampling designs to obtain a random sample of units from within each cluster, such as randomly sampling students in schools (Konold, 2018). Prior research by Lüdtke et al. (2008) and Marsh et al. (2009), for example, has explored measurement issues related to the sampling ratio in multilevel models, with focus on sampling error that arises when individual responses are aggregated to manifest L2 variables. Findings by Lüdtke et al. (2008) suggest that the sampling ratio may be related to bias in contextual effect estimates, in which an aggregated L2 variable has some effect on an outcome after controlling for L1 characteristics. This work was based on a formative aggregation process in which L1 units (e.g., individuals) respond to variables as they personally relate to themselves (e.g., I have been bullied this year). By contrast, reflective L2 constructs are measured through variables in which the referent is at L2 (e.g., ratings obtained by L1 students that are asked about the climate of their L2 school). Moreover, they focused on contextual effect estimates, in which aggregated group-level effects for an outcome are compared with individual-level effects. However, it has yet to be explored how sampling ratios are related to the measurement of L2 latent constructs themselves.

The overarching goal of this paper is to more fully explore the sampling ratio in terms of how it is related to estimates of L2 factor loadings based on latent aggregations of L1 data. Toward that end, we build upon prior research by Lüdtke et al. (2008) who considered the role of the sampling ratio on sampling error in MSEM models. Specifically, we leverage previous

research that has focused on L2 unreliability in contextual effect estimates due to sampling error, which includes both bias and efficiency. To address these gaps, we explored the role of the sampling ratio in MSEM models, by specifically examining the average relative bias and variability in estimates of L2 factor loadings through Monte Carlo simulations.

We begin by first considering how manifest multilevel models can be extended to MSEM for purposes of accounting for sampling error, measurement error, or both. Next, we discuss variability and reliability in MSEM models. We then distinguish between reflective and formative L2 constructs and consider theoretical and the modeling implications. Finally, we consider the interchangeability assumption and its relationship with the sampling ratio and the measurement of L2 latent constructs. After describing the setup of the simulation study, we summarize the various results. Lastly, we provide a general discussion of the findings, offering guidance for applied researchers, while also proposing future directions for additional research.

Multilevel Structural Equation Model (MSEM) Framework

The general MSEM framework proposed by Preacher et al. (2010) integrates CFA/SEM and MLM into a single approach. A particular advantage of MSEM is that it decomposes observed individual ratings into two orthogonal L1 and L2 latent components, allowing for a more nuanced understanding and modeling of error sources. The terms latent components, latent constructs, and latent factors are used interchangeably throughout the text, and are akin to the concept of a latent trait in the context of IRT models. When evaluating L2 group effects through L1 responses, Lüdtke et al. (2011) argue that two types of unreliability may be present: measurement error in the indicators of L1 and L2 constructs, as well as sampling error due to the sampling of L1 individuals within each L2 group that are obtained for the purpose of responding to the manifest variables. They described a 2×2 taxonomy of ML models that account (or fail to

account) for these different types of unreliability that may be present. We outline these four

models below in relation to a two-level structure in which individuals (e.g., students) nested

within groups (e.g., schools) respond to four manifest variables (e.g., individual students

responding to four questions about the climate of their respective schools). Figure 1 provides a

graphic representation of all four models.

The *doubly manifest model (Model 1)* is the most basic model in that it assumes both

measurement and sampling error are zero by failing to explicitly take these sources of variance

into account. Here, the manifest observed variable is decomposed into two orthogonal L1 and L2

components. At the within-level, a single observed composite score $\bar{X}_{ij}$ is calculated by averaging

across the *p* items for each individual *i* in school *j*:

$$\bar{X}_{ij} = \frac{1}{p} * \sum_{p=1}^{P} x_{pij} \tag{1}$$

At the between-level, a single observed composite score $\bar{X}_j$ representing the group (or cluster)

mean is calculated by averaging across the *p* items and the n_j students in each school *j*:

$$\bar{X}_j = \frac{1}{P * n_j} * \sum_{i=1}^{n_j} \sum_{p=1}^{P} x_{pij} \tag{2}$$

Model 1 in Figure 1 provides a graphic representation of this hypothetical model. This model is

doubly manifest in that it combines observed variables to create a composite score (i.e., ignores

measurement error) and that it uses a manifest aggregation from L1 to L2 (i.e., ignores sampling

error). Moreover, this model "may be a highly unreliable measure of the unobserved group

average" when a small number of L1 individuals are sampled from within each L2 group

(Lüdtke et al., 2011, p. 450).

The issue of sampling error arises when individual responses are averaged within L2

units to compute an observed group-mean that is then used to represent L2 constructs (Nagengast

& Marsh, 2011). Because L2 reflective traits are typically assessed through ratings obtained on a

sample of L1 informants (e.g., students or teachers), different random samples may produce

different estimates of the L2 trait. The sampling error inherent in a sample estimate like the mean

can lead to bias in estimating substantive associations between L2 school climate constructs and

other L2 outcomes (Shin & Raudenbush, 2010). This bias can become increasingly pronounced

as the absolute size of the L1 units vary across L2 units. Here, greater weight is given to schools

with a larger number of informants when ratings are aggregated across schools (Wang & Degol,

2016). According to Marsh et al. (2012), sampling error in L2 constructs obtained through

aggregation of L1 response ratings are "a function of the average agreement among individuals

in the same group and the number of sampled individuals in each group" (p.111). When there is

strong agreement on a large number of items among informants within the same school,

estimates of sampling error will be smaller. Consequently, these L1 responses will provide better

estimates of the L2 school trait being measured.

To help control for sampling error, the *manifest-measurement/latent-aggregation model*

(Model 2) separates the manifest observed variable into two orthogonal latent components at L1

and L2 (Muthén, 1989; 1990; 1994) that are then aggregated to obtain the L1 and L2 substantive

trait variables. However, this model is still manifest-measurement in that it assumes no

measurement error through the aggregation of the indicators. In this model, each observed

indicator variable is measured at the within level, and may be decomposed as the sum for each

group plus an individual deviation from the group average (Asparouhov & Muthén, 2007a;

Muthén & Muthén, 2017):

$$X_{ij} = U_{Wij} + U_{Bj} \tag{3}$$

where X_{ij} is the observed response to item x for individual i in school j, U_{Bj} is a latent variable at

the between-level representing the true cluster average, and U_{Wij} is a latent variable at the within-

level representing individual deviation from the cluster average. Thus, a single observed

composite score $\bar{X}_{ij}$ at the within-level is calculated as:

$$\bar{X}_{ij} = \beta_W U_{Wij} + \epsilon_{ij} \tag{4}$$

where β_W represents an estimated regression coefficient, and ϵ_{ij} represents an individual residual

error term. Similarly, a single observed composite score $\bar{X}_j$ at the between-level is calculated as:

$$\bar{X}_j = \mu + \beta_B U_{Bj} + \epsilon_j \tag{5}$$

where μ represents the overall grand mean of X, β_B represents a regression coefficient, and ϵ_j

represents a group-specific residual error term. This is sometimes referred to as a multilevel

latent variable covariate approach (Lüdtke et al., 2008; Marsh et al., 2009). Moreover, this

decomposition of an observed L1 indicator into two uncorrelated L1 and L2 latent variables is

the default setting in Mplus when the variable is not mentioned on the 'within' statement

(Asparouhov & Muthén, 2007a). Model 2 in Figure 1 provides a graphic representation of this

hypothetical model. The main distinction between this model and the doubly manifest model

regards the assumption of an unknown group mean at L2.

Figure 1 also represents a hypothetical *latent-measurement/manifest-aggregation model*

(Model 3), with a single factor structure at both the within and between-level. This model can be

conceived of as an extension of model 1 in that the L1 and L2 trait indicators are manifest, but

substantive traits are formed as latent variables in a manner consistent with confirmatory factor

analysis. This model accounts for measurement error in the indicators as the substantive trait

factors represent shared variance across indicators, and residual sources of variance in the

indicators modeled. The model is latent-measurement in that it takes into account measurement

error at both L1 and L2 by estimating latent substantive factors that are measured by multiple L1

and L2 indicators. However, it is manifest-aggregation in that the observed group mean L2

indicators are the result of manifest aggregations from L1 to L2, thus not taking into account

sampling error.

In the *latent-measurement/manifest-aggregation model*, the L2 latent factor X_j is based on

a manifest aggregation of observed indicators. That is: $X_j = \frac{1}{n_j} * \sum_{i=1}^{n_j} x_{pij}$. Letting X_{pij} represent

a single observed score on item p for person i in group j, the within-level model can be expressed

as:

$$X_{pij} - X_j = \mu_{px} + \lambda_{pW} * U_{xij} + E_{pxij} \tag{6}$$

where μ_{px} represents the overall grand mean of item p; λ_{pW} represents the L1 factor loadings;

and U_{xij} and E_{pxij} represent the unobserved true score and error score at L1, respectively.

Similarly, the between-level model is:

$$X_j = \mu_{px} + \lambda_{pB} * U_{xj} + E_{pxj} \tag{7}$$

where λ_{pB} represents the L2 factor loadings, while U_{xj} and E_{pxj} represent the unobserved true

score and error score at L2, respectively.

Model 4 in Figure 1 represents a hypothetical doubly latent model. This model builds on

the latent-measurement/manifest-aggregation model (Model 3) by not only controlling for

measurement error at L1 and L2, but also controlling for L2 sampling error through the latent L2

indicator aggregation process described for Model 2. In classical test theory, an individual's

observed score X is equal to the sum of the (unobserved) true score U_X plus the error score E_X

(Lord & Novick, 1968). Following Marsh et al. (2009) and Muthén (1991), this can be extended

for multilevel data, where, at the within-level, a single observed score on item p for person i in

group j can be decomposed as:

$$X_{pij} = v_j + \lambda_{pW} * U_{xij} + E_{pxij} \tag{8}$$

while the between-level model is:

$$\nu_j = \mu_{px} + \lambda_{pB} * U_{xj} + E_{pxj} \qquad (9)$$

Equations 8 and 9 may be combined into a single model:

$$X_{pij} = \mu_{px} + \lambda_{pW} * U_{xij} + \lambda_{pB} * U_{xj} + E_{pxij} + E_{pxj} \qquad (10)$$

In contrast to Model 3, the L2 indicator means are no longer manifest observed aggregations

from L1, but rather are considered latent indicator variables with random intercepts. Thus, each

observed indicator in the doubly latent approach is decomposed as is shown in Equations 8 and

9. Identification of this model may be achieved in multiple ways, depending on the central

research question. For instance, to estimate the latent factor variance at both levels, the first

factor loading at both levels may be fixed to one. Conversely, estimation of all factor loadings at

both levels may be achieved by constraining the factor variance at both levels.

All four models can be understood in relation to their corrections for unreliability. Model

1 implicitly assumes there is no measurement error and no sampling error and is thus considered

to be a no correction model. Models 2 and 3 both represent partial correction models, where

Model 2 corrects for sampling error in the aggregation of an L1 construct to form an L2

construct, and Model 3 corrects for measurement error by using multiple observed indicators to

estimate L1 and L2 constructs. The doubly latent model can be considered a full correction

model, in that it simultaneously corrects for both measurement error and sampling error.

ICC and Reliability

A common measure to conceptualize variation in multilevel models is the intra-class

correlation coefficient (ICC). Depending on the MSEM model, there may be both a latent factor

ICC (ICC_L) and an observed variable (ICC_O). In multilevel models without latent constructs, the

ICC_O represents the proportion of an observed variable's variance that can be attributed to the

between-group differences (Raudenbush & Bryk, 2002). More specifically, ICC_O represents the

proportion of total observed score variance due to differences between groups, and can be

defined as:

$$\text{ICC}_\text{O} = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{11}$$

where τ^2 is the variance between groups and σ^2 is the variance within groups. This has been

referred to as ICC(1) in organizational psychology and other disciplines (Bliese, 2000). The ICC_O

is directly related to the reliability of the observed aggregated L2 construct $\boldsymbol{\eta}_{Bj}$ (Bliese, 2000):

$$\text{Reliability}\left(\boldsymbol{\eta}_{\text{Bj}}\right) = \frac{n * \text{ICC}_\text{O}}{1 + (n-1) * \text{ICC}_\text{O}} \tag{12}$$

where n represents the average number of L1 units sampled from group j. The reliability of the

aggregated L2 construct is sometimes referred to as ICC(2) in organizational psychological

literature (Bliese, 2000). Thus, it can be seen that as the ICC_O increases, holding constant group

size, the reliability of the aggregated L2 construct increases. Similarly, holding constant the

ICC_O, as group size increases, so too does reliability.

In MSEM models with latent factors, the ICC_O also takes into account factor loadings,

factor variance, and residual variance at both levels. Let λ_B represent the between-level factor

loading, ψ_B represent the between-level factor variance, and θ_B represent the between-level

residual variance. Similarly, let λ_W represent the within-level factor loading, ψ_W represent the

within-level factor variance, and θ_W represent the within-level residual variance. The ICC_O may

be computed as (Depaoli & Clifton, 2015; Hsu et al., 2016):

$$\text{ICC}_\text{O} = \frac{\left(\lambda_\text{B}^2 * \psi_\text{B} + \theta_\text{B}\right)}{\left(\lambda_\text{B}^2 * \psi_\text{B} + \theta_\text{B}\right) + \left(\lambda_\text{W}^2 * \psi_\text{W} + \theta_\text{W}\right)} \tag{13}$$

Clearly, the ICC_O may have important implications regarding the measurement of aggregated L2

constructs, where larger ICC_O would likely yield more stable measurement of L2 latent

constructs. For MSEM models with latent factors, the ICC_O is directly affected by the factor

loadings, in which stronger factor loadings yield larger within-group correlations, all else equal.

When an identical model structure is assumed at both the within and between levels, researchers

may further calculate the ICC_L:

$$ICC_L \;=\; \frac{\psi_B}{\psi_B + \psi_W} \tag{14}$$

where ψ_B represents the variance of the of the between-level latent factor and ψ_W represents the

variance of the within-level factor (Heck & Kwok, 2012; Hsu et al., 2016).

A second measurement issue concerns sampling errors that arise when averaging

individual responses to form L2 constructs. Equations 8 and 9 provide the multilevel extension

of classical test theory, in which an observed indicator at L1 is decomposed into two uncorrelated

L1 and L2 latent variables, as well as separate error scores at each level. Consequently, reliability

at L2 is defined by two kinds of error: 1) measurement error of E_{pxj} that is due to unreliability of

the observed indicator X_{pij}, and 2) sampling error due to sampling a finite sample from a finite or

infinite population (Lüdtke et al., 2008; Lüdtke et al., 2011). Typically, both L1 and L2

constructs are assessed by obtaining data from a sample of individuals that reside within a L2

population. As a result, different samples of individuals within L2 groups may produce different

estimates of the group-level trait (Jia & Konold, 2019; Nagengast & Marsh, 2011). Moreover, as

noted by Lüdtke et al. (2011, p. 446), "if only a small number of L1 individuals are sampled

from each L2 group, the observed group average may be a highly unreliable measure of the

unobserved true group average U_{xj}." This can be thought of as unreliability in a sample estimate,

and may ultimately lead to bias in estimates between L2 constructs and other L2 outcomes (Shin & Raudenbush, 2010).

Construct Meaning and Implications

Group-level constructs are typically estimated through the collection of information on individuals within each group. In organizational psychology, for example, each individual's rating of job satisfaction may be combined to form an overall group average of job satisfaction within a company. Early work by Cronbach (1976) highlighted potential problems and noteworthy methodological issues that may arise when obtaining group-level estimates through the aggregation of individual observed responses. One issue relates to the referent of L1 ratings, and the nature of the construct under study. Prior work by Lüdtke et al. (2008) and Stapleton et al. (2016), among others, distinguishes between formative and reflective L2 constructs, sometimes referred to in the literature as configural and shared constructs, respectively. The main distinction between the two constructs lies in the reference group, where the individual is the referent in the aggregation process for formative constructs, while the group itself is the referent in the aggregation process for reflective constructs. Moreover, prior research has established various recommendations regarding appropriate modeling of these differing constructs (Jak, 2019; Kim et al., 2016; Stapleton et al., 2016). These considerations are described in detail below.

Reflective Constructs

Consider a process in which student ratings of a common organizational trait like school safety are aggregated to form an overall school-level safety construct. This represents a reflective L2 aggregation process because each student is providing an estimate of a common trait. Observed indicators of reflective constructs are considered isomorphic across units within a

given cluster, where responses to items are seen as interchangeable (or exchangeable), an important concept in multilevel modeling and organizational theory (Bliese, 2000; Bliese et al., 2007; Stapleton et al., 2016). Individuals within an L2 group may be considered interchangeable if their observed indicator responses all have the same relationship to the unobserved group mean U_{xj} (Croon & van Veldhoven, 2007; Lüdtke et al., 2008). Many authors have argued that interchangeability of L1 units within an L2 group holds for the measurement of reflective L2 constructs, in which there is an expectation that the individual level and aggregated variables both reflect the same construct (Croon & van Veldhoven, 2007; Lüdtke et al., 2008; Lüdtke et al., 2011; Marsh et al., 2009). Reflective L2 constructs are assumed to give rise to ratings of L1 observed indicators, analogous with response processes commonly assumed by typical latent variable modeling approaches, in which responses to multiple indicators are used to infer a latent construct. More simply, the responses to L1 indicators are assumed to be influenced by the L2 construct (i.e., causal arrows in the structural model go from the L2 construct to the L1 indicators). Thus, individuals are assumed to have the same relationship to the unobserved group mean. For such scenarios, the ICC_O, an indicator of variation within an L2 group, is used to estimate L2 sampling error due to a finite sampling of individuals.

Further, for reflective L2 constructs, the within-component is often not of interest, as there should in theory be no variability at L1 for a reflective construct. For reflective constructs, Stapleton et al. (2016) have proposed fitting a saturated model at the within-level for reflective constructs. However, Jak (2019) argues that the same result could be achieved by fitting a two-level factor model with factor loadings constrained to be equal across both levels, in which the variance of the L1 factor is assumed to be zero. Jak (2019) further demonstrates the two-level factor model with cross-level invariance to be less parameterized than the saturated model.

Reflective aggregations of L1 constructs may conflate multiple sources of variance, including variability associated with the measurement of the L2 construct, as well as residual variance unique to each individual (Marsh et al., 2012). Consider the example in which student ratings of school safety are aggregated to form an L2 overall school safety construct. It is likely that student ratings of school safety are influenced by both the student's personal experiences of safety as well as their shared experiences and the perceptions generally held among his or her peers. Failing to disaggregate residual variance components in aggregated L2 constructs has been shown to introduce bias in estimates of L2 effects (Lüdtke et al., 2008; Morin et al., 2014).

Formative Constructs

Next, consider a process in which each student's grade point average (GPA) is aggregated to form an overall school-level mean grade point average. This represents a formative L2 aggregation process because the measurement pertains to the student and there is no reason to believe all students should have the same GPA. In contrast to reflective L2 constructs, individuals within an L2 group are not considered to be interchangeable for formative L2 constructs. Rather, they are likely to have different (unique) L1 true scores because they are the referent of their own measurements, as the referent is the individual. More simply, the responses to L1 indicators are assumed to be influenced by the L1 construct (i.e., causal arrows in the structural model go from the L1 construct to the L1 indicators). As a result, there is no expectation that L1 variables and aggregated L2 variables represent the same construct. In contrast to the measurement of reflective constructs, when dealing with formative constructs, the latent factors at both L1 and L2 are often of substantive interest.

To appropriately model formative L2 constructs, researchers have suggested constraining factor loadings to be equal across levels, as the construct "reflects the cluster aggregate of the

individual construct at Level 1" (Stapleton et al., 2016, p. 496). Here, cross-level invariance is

necessary to establish construct validity of formative L2 constructs (Kim et al., 2016).

Importantly, by imposing this cross-level invariance constraint on factor loadings, the covariance

structure at L2 is identified, so long as the factor variance at L1 is constrained. As a result, other

parameters may differ across levels, while the factor variance at L2 remains estimable (Jak,

2019).

The Role of the Sampling Ratio

Assumptions regarding the interchangeability of L1 individuals may have important

implications regarding the reliability of L2 construct estimates. Examining Equation 12, it can be

seen that reliability in the L2 construct estimates does not account for variation at L1. Instead, it

is assumed that L2 unreliability due to sampling error is based on interchangeable individuals. As

average group size (n) increases, all things equal, sampling distributions and standard errors

decrease, while reliability increases (Jia & Konold, 2019). However, the relative size of the

sample of L1 individuals from a population of potential L1 individuals within an L2 group can

play a role when estimating aggregated L2 constructs. Prior simulation work by Lüdtke et al.

(2008) examined the role of the sampling ratio in MSEM models with aggregated formative L2

constructs. The sampling ratio simply represents the proportion of sampled individuals from

within a group:

$$SR_j = \frac{n_j}{N_j} \tag{15}$$

where the sampling ratio for group j (SR_j) is equal to the number of sampled L1 individuals from

within group j (n_j), divided by the total number of L1 individuals in group j (N_j). For example,

sampling 5 students from a class with a total of 25 students would represent a sampling ratio of

20%. Lüdtke et al. (2008) explored both relative percentage bias and root mean square error

(RMSE) in contextual effect estimates for doubly manifest and manifest-measurement/latent-aggregation model (Models 1 and 2 in Figure 1). Results demonstrated that when both the total number of L1 individuals in a group and the sampling ratio are small, both approaches performed poorly. However, when the sampling ratio is low and the sample size is high, the manifest-measurement/latent-aggregation approach outperformed the doubly manifest approach with regards to bias and RMSE. Moreover, with a low sampling ratio (e.g., 5%) and a large number of total L1 individuals per group (e.g., $n = 500$), the finite population sampling model approaches that of an infinite sampling model.

Some (e.g., Marsh et al., 2012) have argued that as the sampling ratio approaches 1.0, it is reasonable to assume no sampling error in the measurement of a L2 construct. Others (e.g., Shin & Raudenbush, 2010) have argued that if each cluster is assumed to be sampled from a larger population of clusters, it is appropriate to account for sampling error by treating the unobserved group means as a latent variable measured with some amount of precision (Asparouhov & Muthén, 2007a). However, the ability to control for sampling error, and thus improve reliability in L2 construct estimates, depends in part on the sampling ratio. As a result, different MSEM models may be better equipped than others to control for the sampling error, leading to more unbiased estimates.

Overall, questions remain about the role of the sampling ratio in MSEM. Much of the previous work on the sampling ratio has considered its impacts on: a) contextual effect estimates in the context of b) measurement Models 1 and 2 in Figure 1 (i.e., doubly manifest and manifest-measurement/latent-aggregation models). However, the effects of the sampling ratio on estimates of L2 factor loadings remains unknown, specifically within a doubly latent MSEM modeling approach (Model 4 in Figure 1). The current study was designed to investigate whether the

sampling ratio is related to bias and variability in aggregated formative L2 construct

measurement and estimation in the context of doubly latent MSEM models.

Methods

Overview of the Analyses

We used Monte Carlo simulation to investigate bias and variability in estimates of factor

loadings for aggregated formative L2 constructs across differing sampling ratios. It is assumed

that the number of L1 units within each L2 group is a finite number (e.g., 100), and that each

cluster is of equal size. A two-step procedure was used to generate populations with finite L1

sample sizes and a fixed number of L2 units. In the first step, clusters were generated to establish

a population model with a finite sample size within each L2 group (e.g., $J = 250$ clusters with N_j

$= 100$ individuals within each L2 cluster). In step two, a sample of L1 units were drawn from

each cluster according to a specified sampling ratio (e.g., 20%). This resulting sample

represented the analytic sample to which a model was fit and estimates were produced. This two-

step procedure was replicated 1,000 times for each condition described below.

Data Generation and Analysis

Mplus version 8.4 (Muthén & Muthén, 2017) was used to generate the data

corresponding to the model presented in Equations 8 and 9, in which a single L1 latent construct

is measured by four observed indicators at the within-level. At the between-level, a single L2

latent construct is measured by four latent aggregated group means corresponding to each

indicator at L1. A graphic representation of this model is provided by Model 4 in Figure 1.

Consistent with Jak (2019) and Stapleton et al. (2016), factor loadings were constrained to be

equal across levels. For each replication, the associated population dataset was saved. Next, for

each population dataset, a random sample of L1 units were drawn from each cluster according to

a specified sampling ratio using R version 4.0.0 (R Core Team, 2020) and saved as an analytic

dataset. Finally, each analytic sample dataset was then analyzed using an external Monte Carlo

simulation study in Mplus. For identification purposes, the variance of the L1 factor was

constrained (to 1) for the model fit to the analytic sample data. This resulted in a total of 17

parameters estimated: four factor loadings (cross-level invariance), four residual variances at L1,

four intercepts at L2, one factor variance at L2, and four residual variances at L2. The

MplusAutomation package (Hallquist & Wiley, 2018) in R was used extensively to facilitate

conducting the simulations in Mplus. Annotated R code used for data generation and analysis is

freely available at https://github.com/jmk7cj/Sampling-Ratio-for-MSEM. An example Mplus

input file used to generate population data is also provided, along with an example Mplus input

file used to analyze the analytic sample datasets.

Simulation Conditions

The following conditions were manipulated: the number of L2 groups ($J = 50, 100, 500$),

the total number of L1 units per L2 cluster ($n_T = 20, 100, 1000$), the ICC_O of the observed L1

indicators ($ICC_O = 0.05, 0.25$), the standardized factor loadings ($\lambda = 0.5$ and 0.8), and the

sampling ratio (SR = 5%, 20%, 50%, 80%). Thus, for example, when the total number of L1

units per L2 cluster was 20 and the sampling ratio was 50%, the number of L1 units sampled

from each cluster was 10. This resulted in a total of $3 \times 3 \times 2 \times 2 \times 4 = 144$ unique simulation

cells. A total of 1,000 datasets were generated and analyzed for each condition.

Values for the conditions considered were informed both by prior MSEM simulation

research and previous applied education research. Previous research by Hox and Maas (2001)

and Maas and Hox (2005) found cluster sizes less than 50 may lead to biased estimates in

multilevel structural equation models. Hox and Maas (2001) argued that more than 100 L2

groups may be needed for unbiased estimates of a between-level model with low ICC_O, as specified below. Similarly, prior MSEM simulation work (Depaoli & Clifton, 2015; Lüdtke et al., 2008) utilized L2 group sizes of 50, 100, and 500. Often in applied educational work, researchers may deal with sample sizes much smaller than 100. As such, we decided to explore a low condition of 50 L2 groups, in addition to 100 and 500 L2 groups.

The total number of L1 observations per L2 cluster were also informed by simulation work from Lüdtke et al. (2008), who considered 25, 100 and 500 L1 observations per L2 cluster. We wanted to extend the larger end of this condition to 1,000 to more closely replicate group sizes that may be encountered in secondary school educational research (e.g., high schools with 1,000 or more students). Regarding ICC_O values, prior educational research and meta-analyses have reported ICC_O values typically ranging from 0.05 to 0.25 (Bloom et al., 2007; Hedges & Hedberg, 2007; Murray & Short, 1995). Similarly, Lüdtke et al. (2008) considered values from .05 to .30 for observed variable ICC_O. Therefore, we chose two different ICC_O: 0.05 and 0.25 within these ranges.

Values for the standardized factor loadings were informed by prior MSEM simulation literature by Hox and Maas (2001), Kim et al. (2012), Lüdtke et al. (2008) and Lüdtke et al. (2011), where values ranged from 0.3 to 0.9. In practice, standardized factor loadings may be even less for MSEM models examining educational data (Jia & Konold, 2019; Morin et al., 2014). Some (e.g., Kline, 2011; Sun et al., 2011) consider standardized loadings ≥ 0.40 as necessary for establishing the validity and reliability of a single indicator. As a result, we chose two values, where standardized factor loadings of 0.5 and 0.8 represent low and high values. These values can be interpreted as reliability of a single indicator, where a value of 0.8 indicates that $0.8^2 \times 100 = 64\%$ of the observed variance can be explained by the latent factor (Lüdtke et

al., 2011). Moreover, it was assumed that the factor loadings were invariant across L1 and L2

(see Appendix for calculations of standardized factor loadings for generating data).

    Lastly, we chose four values of the sampling ratio: 5%, 20%, 50%, and 80%. In the only

other simulation study examining sampling ratios in MSEM models, Lüdtke et al. (2008)

considered sampling ratios of .2, .5, .8, and 1.0. We were interested in extending this to scenarios

in which the sampling ratio is even less than 20%. For example, consider a high school with

1,000 students (as described above), a scenario often faced by educational researchers. A

sampling ratio of 5% would result in a sample of 50 students. Many applied educational

examples deal with less than 50 units per L2 group sampled (Marsh et al., 2012). As a result, we

believe a lower sampling ratio is important in understanding effects when dealing with L2 groups

of larger sizes.

Evaluation Criteria

    All models were estimated in Mplus using maximum likelihood parameter estimation

with robust standard errors (MLR; Muthén & Muthén, 2017). We are most interested in model

estimates of the four standardized factor loadings. We focus on factor loadings for three

important reasons. First, as demonstrated in Equation 13, standardized factor loadings are

directly related to the ICC_O or within-group correlation, a measure often of great interest to

researchers. Second, factor loadings are also used to calculate the reliability of an indicator,

ultimately impacting level-specific reliability (Geldhof et al., 2014). Thus, any errors in

estimated factor loadings may have implications for subsequent measures. Third, factor loadings

themselves are often interpreted as measures of construct validity that convey the degree to

which a latent variable influences responses to items that are presumed to be indicators of the

latent construct. To assess the performance of the model, we focused on three criteria for model

evaluation: 1) the relative percentage bias of the parameter estimate, 2) the accuracy of the

standard error, and 3) the root mean square error (RMSE). The relative percentage bias indicates

the accuracy of the factor loading estimate and is calculated as (Hoogland & Boomsma, 1989;

Lüdtke et al., 2008):

$$B(\hat{\beta}) = 100 * \frac{(\hat{\bar{\beta}} - \beta)}{\beta} \tag{16}$$

where $B(\hat{\beta})$ is the relative percentage bias of the L2 factor loading estimate $\hat{\bar{\beta}}$, as compared to

the known population factor loading β. Following Muthén (2005) and Muthén and Muthén

(2002), we consider relative bias values less than or equal to ±5% as within an acceptable

threshold. To assess the accuracy of the standard error of the standardized factor loading

estimates, we examined parameter coverage (i.e., the proportion of replications in which the 95%

confidence interval contained the true population parameter). Finally, the overall accuracy of the

estimated factor loadings was assessed using the RMSE, where RMSE values are equal to the

square root of the variance of the estimates across replications plus the square of the bias.

To better understand which simulation conditions contributed to bias, parameter

coverage, and RMSE, we conducted four-way factorial analysis of variance (ANOVA) tests, in

which bias, parameter coverage, and RMSE were dependent variables, and each of the

manipulated simulation conditions (L2 sample size, total L1 units per L2 cluster, ICC_O,

standardized factor loading, and sampling ratio) were factors, as well as all interactions among

the factors. To describe the significance of the conditions, omega-squared (ω^2) effect sizes were

calculated for all main effects, as well as two-, three-, and four-way interactions. Thus, the effect

sizes can be interpreted as the proportion of variance in the outcome that can be explained by a

set of factors, controlling for the effects of other independent variables or factors. We focus on

omega-squared values greater than or equal to 0.01 as representing potentially meaningful results, in which small ($\omega^2 = .0099$), medium ($\omega^2 = .0588$), and large ($\omega^2 = .1379$) effect size values correspond to Cohen's (1988) rules of thumb for $d = .2, .5,$ and $.8$, respectively (Albers & Lakens, 2018).

Results

Our three criteria for model evaluation (i.e., bias, parameter coverage, and RMSE) were calculated for each factor loading, then averaged across all four loadings. Estimates for the average relative percentage bias, average 95% parameter coverage, and average RMSE for the four factor loadings are provided in Tables 1 and 2. Several general findings emerged across all conditions. In general, the results appeared to improve (i.e., less bias, greater parameter coverage, and smaller RMSE) as either the number of clusters increased, the number of L1 units within each L2 cluster increased, or the sampling ratio increased. This suggests better measurement of the L2 latent factor as either the L1 or L2 sample size increase, or the sampling ratio increase. All models with the condition of $N = 20$ total L1 units within each L2 cluster and SR = .05 had convergence rates less than 50%. In these scenarios, only a single unit within each L2 cluster was sampled (i.e., 0.05*20), representing a singleton cluster. However, every other condition achieved 100% convergence across the 1,000 replications. A series of factorial ANOVA tests was used to further understand how the simulation facet conditions (i.e., L2 sample size, total L1 units per L2 cluster, ICC_O, standardized factor loading, and sampling ratio) were related to model estimates (see Table 3). Note that those conditions with convergence rates less than 50% were excluded from the ANOVA tests.

Bias

The relationship between certain facets and the relative percentage bias in parameter estimates are shown in Table 3. Main effect tests revealed that variations between L2 sample size, L1 sample size, and the sampling ratio resulted in meaningful ($\omega^2 \geq 0.01$) amounts of variance in the bias of factor loading estimates, with the sampling ratio representing the largest effect ($\omega^2 = 0.094$). Additionally, the four-way interaction of L2 Sample Size × L1 Sample Size × ICC_O × Sampling Ratio was strongly related to variability in bias of factor loading estimates ($\omega^2 = 0.047$). Although these main effect findings reveal general patterns of results for some of the design facets, many of the interactions among facets were also meaningful. In all instances, meaningful four-way interactions engulfed lower-level interactions and main effects, indicating that many of the design facets operated in concert to produce unpalatable results.

To aid in interpretation of these findings, Figure 2 depicts the relationship between the relative percentage bias in estimates of the standardized factor loadings and the sampling ratio, with varying combinations of facets. Examining this figure, it can be seen that the largest values of bias occur for scenarios with a sampling ratio of 0.05. In fact, this sampling ratio of 5% is not plotted in the top panel, in which the total number of L1 units within each L2 cluster is 20, as these models failed to converge. However, it can be seen that bias decreases as the number of clusters increases, the number of L1 units within each L2 cluster increases, and the sampling ratio increases.

While there appears to be considerable variability in bias across the simulation conditions, it is equally important to consider the size of that variability. Overall, values of relative percentage bias were well within the acceptable range of ±5%, with no values outside of ±1% ($M = 0.03$, $SD = 0.14$). This finding indicates virtually no bias in estimated factor loadings

across the conditions, demonstrating a desirable property of maximum likelihood estimation of doubly latent MSEM models.

Parameter Coverage

Next, we considered the proportion of replications in which the 95% confidence interval contained the true population parameter, or parameter coverage. Similar to the results for bias, the main effect of the sampling ratio was related to coverage ($\omega^2 = 0.028$), while the number of clusters sampled was also strongly related, explaining approximately 48% of the variability in parameter coverage ($\omega^2 = 0.480$). The largest effect of three-way interactions occurred for L2 Sample Size $\times$ ICC_O $\times$ Sampling Ratio ($\omega^2 = 0.041$), while the largest four-way interaction effect was for the interaction of L2 Sample Size $\times$ L1 Sample Size $\times$ Std. Loading $\times$ Sampling Ratio ($\omega^2 = 0.027$).

Figure 3 is provided to help interpret these findings, by depicting the relationship between the parameter coverage of estimates of the standardized factor loadings and the sampling ratio, with varying combinations of facets. Exploring this figure, it can be seen that parameter coverage improves toward the correct 0.95 level as the number of L2 clusters increases. Similarly, as the sampling ratio increases, parameter coverage tends to increase as well. Following the findings from Figure 3, scenarios in which the total number of L1 units within each L2 cluster is 20 and the sampling ratio is 0.05 are not provided in the top panel, as these models failed to converge.

Again, it is necessary to simultaneously consider the relative size of the variability of parameter coverage across simulation conditions. Overall, parameter coverage values were extremely close to the correct 0.95 value, with values ranging from 0.93 to 0.96 (*M* = 0.945, *SD* = 0.01). Thus, the standard errors for point estimates of factor loadings in doubly latent models

demonstrate appropriate 95% confidence intervals almost perfectly match the appropriate 95%

confidence intervals. While the sampling ratio is related to variability in parameter coverage

across simulation conditions, the overall magnitude of these differences is essentially ignorable.

RMSE

An overall measure of accuracy was captured by calculating the RMSE. All main effects

excluding the ICC_O were meaningfully related to the RMSE of estimated standardized factor

loadings, with the sampling ratio accounting for approximately 16% of the variability alone.

Various two-way and three-way interactions had strong effect size estimates, including the two-

way interaction of L1 Sample Size × Sampling Ratio ($\omega^2 = 0.027$). Again, to aid in the

interpretation of these findings, Figure 4 provides a graphical representation of the relationship

between the RMSE of estimated standardized factor loadings and the sampling ratio, with

varying combinations of facets. Figure 4 demonstrates that accuracy in the average estimates of

factor loadings increases (i.e., RMSE decreases toward zero) as the number of clusters increases,

the number of L1 units within each L2 cluster increases, and the sampling ratio increases. Again,

scenarios in which the total number of L1 units within each L2 cluster is 20 and the sampling

ratio is 0.05 are not provided in the top panel, as these models failed to converge.

Finally, although there is substantial variability in the RMSE of factor loading estimates,

the overall size of that variability must also be considered. Examining Figure 4, it can be seen

that RMSE values are relatively small, with all but one condition resulting in RMSE values

below 0.20 ($M = 0.04$, $SD = 0.04$). Further analyses indicated RMSE nearly perfectly matched

the parameter standard error ($r = 0.997$), where RMSE values were a function of the L1 sample

size (Snijders & Bosker, 1993). These findings are consistent with doubly latent estimation

reported by Lüdtke et al. (2011). As a result, although the variability in RMSE differs for various

simulation facets, the maximum likelihood estimates produced by the doubly latent yielded

reliably accurate results.

For Applied Researchers: Illustrative Example

In this section, we considered a case example using empirical survey data from the Early

Childhood Longitudinal Study of Kindergarten (ECLS-K) study to illustrate the application and

interpretation of doubly latent models. The ECLS-K study followed children from the

kindergarten class of 1998-99 through the spring of 8th grade in 2007, focusing on individual and

school-level factors associated with early school experiences and performance (Tourangeau et

al., 2009). For our example, we examined student achievement scores using public-release data

in which students were nested within schools. While the original study involved a three-stage

stratified sampling design, we assumed a simple random sample of schools and students within

each school for simplicity. Additionally, listwise deletion was used to treat missing data. We

restricted our sample to students with achievement scores measured in the spring of 5th grade

(2004). This resulted in a sample of 10,447 students nested within 2,103 schools. The number of

students sampled per school ranged from 1 to 34, with an average of 5 students per school.

Moreover, based on school total enrollment, this resulted in a sampling ratio ranging from 0.13%

to 18.7%, with an average sampling ratio of 3.7%. Thus, for example, if a school had a total

enrollment of 500 students and 25 students were sampled, the sampling ratio was 5%.

We explored the possibility of a single latent factor at L1 (students) underlying three

continuous variables representing student achievement scores in reading, math, and science,

while also aggregating to a single latent factor at L2 (schools). This represents a formative L2

aggregation process as the measurement pertains to the student, and there is no reason to believe

all students should have the same values for the three achievement variables. For formative

constructs, students are not considered to be interchangeable, but rather are likely to have

different, unique L1 true scores. In contrast to the measurement of reflective constructs (i.e.,

where the target of measurement is at a higher level), the latent factors at both L1 and L2 are

often of substantive interest with formative constructs. Thus, a L1 indicator may represent a

measure of student achievement, while a L2 indicator may represent the average student

achievement within a school.

The three L1 achievement variables ranged from 0 to 96, in which reading ($M = 50.28$,

$SD = 10.13$), math ($M = 50.68$, $SD = 9.95$), and science ($M = 50.37$, $SD = 10.10$) were all

approximately normally distributed. Following the methods outlined in the simulation study, a

doubly latent model with cross-level invariance was estimated in Mplus using maximum

likelihood parameter estimation with robust standard errors. By fixing the L1 factor variance to

one, a total of 13 parameters were estimated: three factor loadings (cross-level invariance), three

residual variances at L1, three intercepts at L2, one factor variance at L2, and three residual

variances at L2. We also estimated the model using the lavaan package (Rosseel, 2012) in R,

although we do not report the results here as the output of the two programs was virtually

identical. The dataset and annotated syntax for Mplus and lavaan are freely available at

https://github.com/jmk7cj/Sampling-Ratio-for-MSEM.

Overall, the model fit the data well, with model fit statistics of CFI = .997, TLI = .991,

RMSEA = .045, and SRMR = .015. Estimated intraclass correlations at L1 for reading, math, and

science achievement scores were 0.28, 0.26, and 0.35, respectively; values that are typical for

educational achievement data in elementary school-aged students (Hedges & Hedberg, 2007).

The ICC_O value for reading achievement, for example, can be interpreted as follows:

"Approximately 28% of the variance in reading achievement scores can be attributed to

differences between schools." Moreover, reliabilities of the aggregated L2 components,

commonly referred to as the ICC(2) (Bliese, 2000; see Equation 12), were relatively strong for

reading (ICC(2) = 0.66), math (ICC(2) = 0.63), and science (ICC(2) = 0.72), indicating

satisfactory values (Morin et al., 2014). However, a larger average cluster size (and thus a larger

sampling ratio) would improve ICC(2) reliability estimates.

Table 4 presents the unstandardized parameter estimates. All three factor loadings were

significantly related to students' individual achievement at L1, as well as with average student

achievement at L2, supporting the appropriateness of the measurement model. The estimated

variance of the latent achievement factor at L2 was 0.56, while the variance of the latent

achievement factor at L1 was constrained to one to allow for estimates of all factor loadings.

However, imposing cross-level invariance ensures the latent factors are on a common scale,

allowing for a direct comparison of latent factor variances across levels (Mehta & Neale, 2005).

The proportion of variance explained by the latent achievement factor at L1 ranged from $R^2 =$

0.64 to 0.68, while values at L2 ranged from $R^2 = 0.84$ to 0.98.

Discussion

Multilevel models are often used by social science researchers to estimate effects of

constructs at different levels on outcomes of interest. Various models have been proposed to

control for measurement error and sampling error when measuring latent constructs at L2

through the aggregation of observed indicators at L1. Prior methodological and simulation

research has developed the doubly latent approach that corrects for measurement error and

sampling error, resulting in unbiased estimates of L2 constructs under certain conditions (Lüdtke

et al., 2011; Marsh et al., 2009; Muthén, 1990). This present study adds to the literature by

considering the role of the sampling ratio in doubly latent measurement models.

Our findings suggested that research designs utilizing the doubly latent model with low sampling ratios may face convergence problems. Similar problems with were demonstrated by Lüdtke et al. (2011) who found that doubly latent models often failed to converge for conditions with low ICC_O, and small sample sizes at L1 and L2. Specifically, the convergence rates were severely low for conditions with a L1 sample size of 20 and a sampling ratio of 5%. As previously noted, this condition represents a singleton cluster, in which a single unit is sampled from within each cluster, and estimates of factor loadings are based on information from a sole respondent. While prior research by Bell et al. (2008; 2009) and Clarke and Wheaton (2007) found that both bias and interval estimates were larger for designs in which singleton clusters were present (e.g., up to 70% of clusters were singletons), the problem encountered in the current study is distinctly different. For the conditions with L1 sample size of 20 and sampling ratio of 5%, *all* clusters sampled were singleton clusters, resulting in under-identified models. Thus, the proportion of singleton clusters must be under 100% for model identification purposes. For such scenarios with at least some singleton clusters present, incorporating prior information into the estimation procedure (i.e., Bayesian framework; Gelman et al., 2004) and certain regularization methods (e.g., Yuan & Chan, 2008) to ensure a positive definite covariance matrix, may improve the stability of estimates from models with small sample sizes. With the appropriate specification of informative priors (e.g., half-Cauchy), Bayesian estimation may be well-suited for MSEM with small sample sizes (see Gelman, 2006; McNeish, 2016). As such, this would be a worthwhile avenue for future methodological studies.

Results from the ANOVA indicated that as a main effect, lower sampling ratios also had negative effects on bias, coverage, and RMSE of estimated factor loadings. These findings appear to suggest that even after controlling for other design facets such as L1 and L2 sample

size or ICC_O, the sampling ratio is related to the quality of MSEM model estimates. However, a closer examination of the overall magnitude of the variability revealed a different takeaway. While lower sampling ratios did produce larger bias, all values of bias were within ±1% ($M = 0.03$, $SD = 0.14$), well within the ±5% acceptable range. Similarly, parameter coverage values almost perfectly matched the correct 0.95 value, ranging from 0.93 to 0.96 ($M = 0.945$, $SD = 0.01$). Finally, although smaller sampling ratios were directly related to larger RMSE, the size of the RMSE was relatively small, with all but one condition resulting in RMSE values below 0.20 ($M = 0.04$, $SD = 0.04$). Taken together, these findings demonstrate desirable properties of maximum likelihood estimation of doubly latent MSEM models. Overall, there was virtually no bias in estimated factor loadings across the conditions. The standard errors for point estimates of factor loadings appropriately matched 95% confidence interval coverage, and RMSE values were extremely close to zero, indicating reliable, accurate estimates. As a result, the doubly latent model does appear to appropriately account for both sampling error and measurement error.

Among the various factors investigated in this study, we highlight the importance of the number of clusters sampled and the sampling ratio; these two factors are ones that researchers are likely to have the most control over, whereas it may be more difficult to control other facets such as the ICC_O or factor loadings. Moreover, research has demonstrated that increasing the number of clusters sampled and the number of units per cluster sampled improves statistical power in multilevel models, although a larger gain in power is achieved with larger numbers of clusters (Snijders & Bosker, 1993; Spybrook et al., 2011). Because of this, we encourage researchers to consider increasing both the number of clusters sampled and the sampling ratio to decrease bias and variation in estimated factor loadings. However, it should be noted that incorporating larger samples at L2 into a research design often comes at a greater financial cost

than sampling a larger number of L1 units from within L2 clusters. For an additional discussion,
see Raudenbush and Liu (2000) among others, who consider optimal costs and efficient
allocation of resources in multilevel designs.

In applied research studies, such as students nested within schools, the number of L1
units per cluster can vary considerably in size. For example, Larson et al. (2020) examined the
effectiveness of classroom management practices on student engagement in secondary schools,
in which $N = 54$ high schools ranged in total enrollment size from 323 students to 2,021 students.
Consider an example research design in which a fixed sample of 50 students per school was
collected. This could result in a sampling ratio ranging from approximately 2.5% to 15.5%,
depending on the overall enrollment of the school. While findings from our simulation indicated
that smaller sampling ratios (i.e., 5%) produced larger bias, worse parameter coverage, and larger
RMSE, the overall size of these errors was essentially negligible. Thus, study designs with
smaller sampling ratios of the type investigated here can be used to utilize doubly latent MSEM
models for trustworthy results.

It is also important to note some software capabilities and limitations. For example, the
doubly latent multilevel modeling approach, in which an individual's observed score is
simultaneously decomposed into two uncorrelated L1 and L2 latent variables plus separate error
scores at each level, is the default setting in Mplus (Asparouhov & Muthén, 2007a; Muthén &
Muthén, 2017). As first outlined by Lüdtke et al. (2011) and described in detail above, the
doubly latent model is the only MSEM that corrects for both measurement error and sampling
error. This is also the default estimation procedure for the lavaan package in R (Rosseel, 2012),
which produced nearly identical parameter estimates and standard errors at both levels to those
we obtained from Mplus. Likewise, the generalized linear latent and mixed model (gllamm)

command in Stata 16 (Rabe-Hesketh et al., 2004; StataCorp, 2019) is also capable of estimating

such models, while the gsem command is also capable in theory. However, more traditional

multilevel software such as HLM 8 (Raudenbush et al., 2019) employ an observed variable

disaggregation process to separate between- and within-group effects (Preacher et al., 2010),

leading to an inability to estimate parameters and standard errors at each level simultaneously. In

summary, the estimation techniques and capabilities of various software must always be

considered by researchers and future studies.

Limitations and Future Directions

There are a number of limitations to our simulation study that should be considered, such

as our use of multivariate normally distributed data. One assumption of maximum likelihood as a

normal theory estimator is multivariate normality (Bollen, 1989). However, in educational and

social sciences it is common for researchers to collect dichotomous and ordinal scale data.

Different versions of weighted least squares estimators have been proposed to appropriately

analyze multilevel models for categorical variables (Asparouhov & Muthén, 2007b; Muthén,

1984). However, all models in the current study used maximum likelihood parameter estimation

with standard errors that are robust to nonnormality based on corrections by Yuan and Bentler

(2000). While all of the indicators and latent factors at both levels were drawn from standard

normal distributions in this study, MLR estimation in Mplus, as well as weighted least squares

estimators, may protect against violations of multivariate normality. However, findings from this

study should not be generalized to designs with dichotomous or count indicator variables.

Instead, future researchers should consider the role of estimators specifically designed to handle

categorical variables in MSEM and doubly latent models.

Second, the data generated for all conditions in our simulations were produced from a

doubly latent true population model. Thus, the population model and the analytic model had the

same factor structure at both levels and the same number of observed items at L1, albeit with

differing sample sizes. However, it is possible and often useful to analyze data using different

specifications than those used to generate the population model. Purposefully analyzing data

with a model using the incorrect factor structure, using modeling approaches other than the

doubly latent model, or selecting a sample of items from which to estimate factors may all be

areas to investigate for future simulation research.

Third, the models we examined were assumed to represent formative L2 constructs, in

which individuals within a L2 group were considered to be structurally different (and not

interchangeable). At the same time, evaluating the sampling ratio for reflective L2 constructs

may be impractical, as L1 units in this context are assumed to be interchangeable and thus have

the same relationship to the unobserved group mean (Croon & van Veldhoven, 2007; Marsh et

al., 2009). As a result, implications of the sampling ratio are most apt for formative L2

constructs.

Finally, we utilized a simple random sample of L1 units from each L2 cluster in our

simulation study. As a result, for each replication, each unit within cluster j had a positive, equal

probability of being selected for the analytic sample. However, it is possible to consider the

effects of an unequal probability sampling design, in which the probability of being sampled is

related to some known variable (e.g., only students with high test scores are sampled). Here,

models would need to take into consideration the non-random differences among units and how

these differences may be related to estimates of constructs of both L1 and L2. Again, future

studies could construct an unequal probability sampling procedure relating to the sampling ratio

to some other variable to investigate bias in parameter estimates. Further, the use of quasi-experimental techniques (e.g., propensity scores) may play a role in balancing the covariate distribution, leading to improved estimates.

Conclusions and Implications

Multilevel structural equation modeling continues to make significant progress, with in-depth investigations of methodological techniques. The current study adds to the literature by exploring the role of the sampling ratio in MSEM models with a focus on the doubly latent model. The findings from our simulations indicated that while lower sampling ratios were related to increased bias, increased standard errors, and increased RMSE, the overall size of these errors was negligible, making the doubly latent model an appealing choice for researchers. The doubly latent model was originally proposed as an alternative to more simple ML models, with the ability to account for both sampling error and measurement error. Findings from our study demonstrated that estimation of L2 factor loadings in doubly latent models of formative L2 constructs produced accurate, reliable results across simulation conditions, even with varying sampling ratios, a property that researchers have long expected to be true. These findings have broad implications for educational, psychological, and social science research more generally, in which individuals are clustered within groups, and both L1 and L2 latent constructs are of interest. Future researchers are encouraged to utilize the doubly latent MSEM model for designs with smaller sampling ratios, as the model allows for the decomposition of a single indicator variable into within- and between-level specific components while correcting both sampling error and measurement error.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate

    effect size estimators and follow-up bias. *Journal of Experimental Sociol Psychology, 74*,

    187-195. https://doi.org/10.1016/j.jesp.2017.09.004

Asparouhov, T., & Muthén, B. (2007a). *Constructing covariates in multilevel regression* (Mplus

    Web Notes No. 11, Version 2). Retrieved from:

    http://www.statmodel.com/download/webnotes/webnote11.pdf

Asparouhov, T., & Muthén, B. O. (2007b). *Computationally efficient estimation of multilevel*

    high-dimensional latent variable models. Paper presented at the Joint Statistical Meeting,

    Salt Lake City, UT. Retrieved from: https://www.statmodel.com/papers.shtml

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster size in multilevel models: The impact

    of sparse data structures on point and interval estimates in two-level models. *Proceedings*

    of the joint statistical meetings, survey research section (p. 1122-1129). Alexandria, VA:

    American Statistical Association.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2009). The effect of sparse data structures and

    model misspecification on point and interval estimates in multilevel models. Annual

    Meeting of the American Educational Research Association, San Diego, CA.

Bentler, P. M., & Chou, C. (1987). Practical issues in structural modeling. *Sociological Methods*

    & Research, 16(1), 78-117. doi: 10.1177/0049124187016001004

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for

    data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel*

    theory, research, and methods in organizations (pp. 349–381). Jossey-Bass

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in

    measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research*

    Methods, 10(4), 551-563. doi: 10.1177/1094428107301102

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve

    precision: Empirical guidance for studies that randomize schools to measure the impacts

    of educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

    https://doi.org/10.3102/0162373707299550

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research

    using cluster analysis to create synthetic neighborhoods. *Sociological Methods &*

    Research, 35(3), 311-351. https://doi.org/10.1177/0049124106292362

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York:

    Academic Press.

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design,*

    and analysis. Stanford, CA: Stanford University Evaluation Consortium

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables

    from variables measured at the individual level: A latent variable multilevel model.

    Psychological Methods, 12(1), 45-57. doi: 10.1037/1082-989X.12.1.45

Depaoli, S., & Clifton, J. P. (2015). A bayesian approach to multilevel structural equation

    modeling with continuous and dichotomous outcomes. *Structural Equation Modeling, A*

    Multidisciplinary Journal, 22(3), 327-351. doi: 10.1080/10705511.2014.937849

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

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*(3), 515–534. doi:10.1214/06-BA117A

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25*(4), 621-638. doi: 10.1080/10705511.2017.1402334

Hedges, L., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60-87. https://doi.org/10.3102/0162373707299706

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*(2), 157-174. doi: 10.1207/S15328007SEM0802_1

Hsu, H., Lin, J., Kwok, O., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: A monte carlo study. *Educational and Psychological Measurement, 77*(1), 5-31. https://doi.org/10.1177/0013164416642823

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling, 26*(4), 607-622. https://doi.org/10.1080/10705511.2018.1534205

Jia, Y., & Konold, T. (2019). Moving to the next level: Doubly latent multilevel mediation models with a school climate illustration. *The Journal of Experimental Education*. https://doi.org/10.1080/00220973.2019.1675136

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

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York: Guilford Press.

Konold, T.R. (2018). A multilevel MTMM approach to estimating the influences of contextual factors on trait and informant-based method effects in assessment of school climate. *Journal of Psychoeducational Assessment, 36*(5), 464-476. doi: 10.1177/0734282916683286

Larson, K. E., Pas, E. T., Bottiani, J. H., Kush, J. M., & Bradshaw, C. P. (2020). A multidimensional and multilevel examination of student engagement and secondary schools teachers' use of classroom management practices. *Journal of Positive Behavior Interventions,* 1-14. https://doi.org/10.1177/1098300720929352

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*(4), 444-467. doi: 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. doi: 10.1037/a0012869

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. Methodology, 1(3), 86-92. doi: 10.1027/1614-2241.1.3.86

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin,A. J. S., 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. doi: 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. doi:10.1080/00273170903333665

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling, 23*(5), 750-773. https://doi.org/10.1080/10705511.2016.1186549

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

Morin, A. J. S., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*(2), 143–167. doi:10.1080/00220973.2013.769412

Murray, D. M., & Short, B. (1995). Intra-class correlation among measures related to alcohol use by young adults: Estimates, correlates, and applications in intervention studies. *Journal of Studies on Alcohol, 56*(6), 681-692. doi: 10.15288/jsa.1995.56.681

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered

    categorical, and continuous latent variable indicators. *Psychometrika, 49*(1), 115-132.

    https://doi.org/10.1007/BF02294210

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika,*

    54(4), 557-585. doi: 10.1007/BF02296397

Muthén, B. O. (1990). Mean and covariance structure analysis of hierarchical data. Paper

    presented at the Psychometric Society meeting in Princeton, NJ, June 1990. UCLA

    Statistics Series 62. Retrieved from:

    http://www.statmodel.com/bmuthen/articles/Article_032.pdf

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods &*

    Research, 22(3), 376-398. doi: 10.1177/0049124194022003006

Muthén, B. O. (2005). *Monte Carlo analysis*. Retrieved September 23, 2020, from

    http://statmodel.com/discussion/messages/11/659.html?

Muthén, B., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel

    analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds),

    Handbook of advanced multilevel analysis (pp. 15-40). Taylor and Francis.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample

    size and determine power. *Structural Equation Modeling, 9*(4), 599-620.

    https://doi.org/10.1207/S15328007SEM0904_8

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

    Muthén & Muthén

Nagengast, B., & Marsh, H. W. (2011). The negative effect of school-average ability on science

    self-concept in the UK, the UK countries and around the world: The Big-Fish-Little-

Pond-Effect for PISA 2006. *Educational Psychology, 31*(5), 629-656.

https://doi.org/10.1080/01443410.2011.586416

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

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rabe-Hesketh, S. & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata* (2nd ed.). College Station, TX: StataCorp LP

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*(2), 167–190. doi: 10.1007/BF02295939

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2019). *HLM 8: Hierarchical Linear and Nonlinear Modeling*. Skokie, IL: Scientific Software International, Inc.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199-213. https://doi.org/10.1037/1082-989X.5.2.199

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2). doi: 10.18637/jss.v048.i02

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35*(1), 26–53. doi:10.3102/1076998609345252

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237-259. doi: 10.2307/1165134

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.

Spybrook, J., Bloom, H. Cogdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" software. Available at: http://hlmsoft.net/od/od-manual-20111016-v300.pdf

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. doi: 10.3102/1076998616646200

StataCorp. (2019). *Stata: Release 16*. College Station, TX: StataCorp LLC.

Sun, S., Konold, T. R., & Fan, X. (2011). Effects of latent variable nonnormality and model misspecification on testing structural equation modeling interactions. *The Journal of Experimental Education, 79*(3), 231-256. doi: 10.1080/00220973.2010.481683

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). *Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K-8 full sample data files and electronic codebooks*. Washington, DC: National Center for Education Statistics.

Wang, M.-T., & Degol, J. L. (2016). School climate: A review of the construct, measurement, and impact on student outcomes. *Educational Psychology Review, 28*(2), 315-352. doi: 10.1007/s10648-015-9319-1

Yuan, K. H. & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological Methodology 2000* (pp. 165-200). Boston, MA: Basil Blackwell.

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. Computational Statistics and Data Analysis, 52, 4842–4858. https://doi.org/10.1016/j.csda.2008.03.030

Table 1

Average Relative Bias, Parameter Coverage, and RMSE for Standardized Loading = 0.5

| | $ICC_O = .05$ | | | | | | | | |
| | $N = 20$ *Total L1 units within each L2 cluster* | | | $N = 100$ *Total L1 units within each L2 cluster* | | | $N = 1{,}000$ *Total L1 units within each L2 cluster* | | |
	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE
J = 50									
SR = .05	-	-	-	-0.545	0.947	0.179	-0.060	0.944	0.055
SR = .20	-0.675	0.950	0.199	0.228	0.936	0.088	0.003	0.940	0.028
SR = .50	0.083	0.945	0.124	0.115	0.940	0.055	0.005	0.945	0.017
SR = .80	-0.130	0.936	0.100	0.013	0.940	0.044	0.005	0.939	0.014
J = 100									
SR = .05	-	-	-	-0.128	0.950	0.124	0.018	0.943	0.038
SR = .20	-0.280	0.950	0.142	0.035	0.948	0.061	0.010	0.945	0.019
SR = .50	0.173	0.939	0.088	-0.035	0.949	0.038	0.005	0.946	0.011
SR = .80	0.030	0.940	0.070	-0.013	0.944	0.030	0.003	0.947	0.010
J = 500									
SR = .05	-	-	-	0.028	0.950	0.055	0.000	0.947	0.017
SR = .20	-0.155	0.951	0.063	-0.045	0.952	0.027	0.008	0.949	0.010
SR = .50	0.000	0.953	0.038	0.025	0.950	0.017	0.000	0.947	0.000
SR = .80	-0.025	0.954	0.030	0.003	0.951	0.014	0.003	0.951	0.000
	$ICC_O = .25$								
J = 50									
SR = .05	-	-	-	-0.173	0.946	0.182	-0.098	0.943	0.054
SR = .20	0.330	0.942	0.204	0.058	0.937	0.087	0.005	0.946	0.027
SR = .50	-0.188	0.942	0.124	0.028	0.938	0.055	0.038	0.944	0.017
SR = .80	0.040	0.928	0.102	0.110	0.944	0.043	0.000	0.945	0.014
J = 100									
SR = .05	-	-	-	0.073	0.949	0.127	-0.008	0.949	0.037
SR = .20	-0.070	0.941	0.145	0.023	0.944	0.061	0.040	0.948	0.019
SR = .50	0.218	0.939	0.088	-0.025	0.954	0.038	0.005	0.947	0.011
SR = .80	0.068	0.936	0.071	-0.058	0.942	0.030	-0.010	0.945	0.010
J = 500									
SR = .05	-	-	-	-0.170	0.949	0.056	0.008	0.954	0.017
SR = .20	-0.065	0.954	0.063	0.008	0.951	0.027	-0.010	0.955	0.010
SR = .50	-0.035	0.954	0.038	-0.008	0.950	0.017	-0.010	0.947	0.000
SR = .80	-0.038	0.951	0.030	0.008	0.948	0.014	0.000	0.947	0.000

Note: The conditions with SR = .05 and *N* = 20 total L1 units within each L2 cluster had convergence rates less than 50%.

Table 2
Average Relative Bias, Parameter Coverage, and RMSE for Standardized Loading = 0.8

	$ICC_O= .05$								
	N = 20 *Total L1 units within each L2 cluster*			*N* = 100 *Total L1 units within each L2 cluster*			*N* = 1,000 *Total L1 units within each L2 cluster*		
	Bias	Coverage	RMSE	Bias	Coverage	RMSE	Bias	Coverage	RMSE
J = 50									
SR = .05	-	-	-	-0.623	0.942	0.072	-0.033	0.944	0.022
SR = .20	-0.760	0.946	0.081	0.015	0.945	0.034	-0.013	0.939	0.010
SR = .50	-0.188	0.931	0.052	-0.025	0.939	0.022	0.003	0.942	0.000
SR = .80	0.015	0.939	0.041	0.005	0.941	0.017	0.003	0.945	0.000
J = 100									
SR = .05	-	-	-	-0.035	0.950	0.051	0.018	0.948	0.014
SR = .20	-0.153	0.950	0.057	-0.023	0.948	0.025	0.013	0.945	0.010
SR = .50	0.133	0.944	0.036	0.008	0.944	0.015	0.005	0.947	0.000
SR = .80	-0.028	0.943	0.028	-0.018	0.948	0.011	0.005	0.944	0.000
J = 500									
SR = .05	-	-	-	-0.063	0.947	0.023	-0.018	0.955	0.000
SR = .20	-0.025	0.945	0.027	0.020	0.954	0.010	0.003	0.952	0.000
SR = .50	-0.073	0.943	0.017	-0.015	0.953	0.000	0.003	0.950	0.000
SR = .80	-0.033	0.943	0.014	0.000	0.952	0.000	0.000	0.944	0.000
	$ICC_O= .25$								
J = 50									
SR = .05	-	-	-	-0.395	0.937	0.074	-0.048	0.938	0.022
SR = .20	-0.298	0.937	0.085	-0.165	0.938	0.036	-0.003	0.939	0.010
SR = .50	-0.015	0.934	0.052	-0.028	0.937	0.022	-0.025	0.939	0.000
SR = .80	0.033	0.934	0.041	0.020	0.942	0.017	0.005	0.943	0.000
J = 100									
SR = .05	-	-	-	-0.125	0.945	0.052	0.033	0.940	0.016
SR = .20	-0.108	0.942	0.060	0.010	0.949	0.025	0.010	0.952	0.010
SR = .50	-0.005	0.945	0.036	-0.028	0.947	0.014	0.005	0.939	0.000
SR = .80	-0.020	0.939	0.028	-0.033	0.947	0.011	-0.005	0.944	0.000
J = 500									
SR = .05	-	-	-	-0.005	0.952	0.022	-0.015	0.950	0.003
SR = .20	0.050	0.946	0.027	-0.018	0.951	0.010	-0.008	0.954	0.000
SR = .50	-0.075	0.949	0.017	-0.005	0.946	0.005	-0.005	0.945	0.000
SR = .80	-0.038	0.945	0.014	0.000	0.955	0.000	-0.003	0.954	0.000

Note: The conditions with SR = .05 and *N* = 20 total L1 units within each L2 cluster had convergence rates less than 50%.

Table 3

Omega-Squared Values for Analysis of Variance Effects of Simulation Conditions

	Bias	Coverage	RMSE
Main Effects			
L2 Sample Size	**0.042**	**0.480**	**0.155**
L1 Sample Size	**0.026**	**0.046**	**0.278**
ICC_O	0.008	0.005	0.000
Std. Loading	0.008	0.007	**0.158**
Sampling Ratio	**0.094**	**0.028**	**0.161**
2-way Interactions			
L2 Sample Size × L1 Sample Size	**0.011**	**0.018**	**0.044**
L2 Sample Size × ICC_O	**0.012**	0.008	0.000
L2 Sample Size × Std. Loading	**0.021**	**0.010**	0.000
L2 Sample Size × Sampling Ratio	0.010	0.006	**0.027**
L1 Sample Size × ICC_O	0.000	0.009	**0.040**
L1 Sample Size × Std. Loading	-0.001	0.002	0.000
L1 Sample Size × Sampling Ratio	**0.094**	0.002	**0.027**
ICC_O × Std. Loading	**0.107**	**0.020**	**0.051**
ICC_O × Sampling Ratio	**0.018**	-0.002	0.000
Std. Loading × Sampling Ratio	-0.006	**0.019**	**0.025**
3-way Interactions			
L2 Sample Size × L1 Sample Size × ICC_O	**0.015**	0.009	0.000
L2 Sample Size × L1 Sample Size × Std. Loading	-0.001	**0.037**	0.008
L2 Sample Size × L1 Sample Size × Sampling Ratio	-0.003	0.001	0.000
L2 Sample Size × ICC_O × Std. Loading	-0.004	**0.010**	0.000
L2 Sample Size × ICC_O × Sampling Ratio	**0.063**	**0.041**	0.009
L2 Sample Size × Std. Loading × Sampling Ratio	0.004	0.003	0.000
L1 Sample Size × ICC_O × Std. Loading	**0.053**	**0.030**	0.000
L1 Sample Size × ICC_O × Sampling Ratio	0.007	0.004	0.004
L1 Sample Size × Std. Loading × Sampling Ratio	-0.013	0.005	**0.010**
ICC_O × Std. Loading × Sampling Ratio	-0.002	0.007	0.000
4-way Interactions			
L2 Sample Size × L1 Sample Size × ICC_O × Std. Loading	-0.010	-0.002	0.000
L2 Sample Size × L1 Sample Size × ICC_O × Sampling Ratio	**0.047**	0.001	0.000
L2 Sample Size × L1 Sample Size × Std. Loading × Sampling Ratio	-0.003	**0.027**	0.004
L2 Sample Size × ICC_O × Std. Loading × Sampling Ratio	-0.001	**0.015**	0.000
L1 Sample Size × ICC_O × Std. Loading × Sampling Ratio	-0.004	0.005	0.000

Note: Omega-squared effect sizes of small ($\omega^2 > .0099$), medium ($\omega^2 > .0588$), and large ($\omega^2 > .1379$).

Table 4

Illustrative Example Results: Unstandardized
Parameter Estimates of Student Achievement

	Estimate	*SE*
Level-1 (within)		
Factor loadings		
Reading	6.83	0.083
Math	6.63	0.083
Science	6.68	0.076
Variance		
Reading residual	22.01	0.618
Math residual	24.38	0.564
Science residual	21.91	0.564
Achievement factor	1.00	NA
R-square		
Reading	0.68	0.008
Math	0.64	0.009
Science	0.67	0.008
Level-2 (between)		
Factor loadings		
Reading	6.83	0.083
Math	6.63	0.083
Science	6.68	0.076
Intercepts		
Reading	50.30	0.161
Math	50.46	0.157
Science	50.05	0.174
Variance		
Reading residual	0.42	0.308
Math residual	2.68	0.395
Science residual	4.48	0.425
Achievement factor	0.56	0.037
R-square		
Reading	0.98	0.011
Math	0.90	0.014
Science	0.84	0.015

Note: The variance of the latent achievement factor
at L1 was constrained to one.

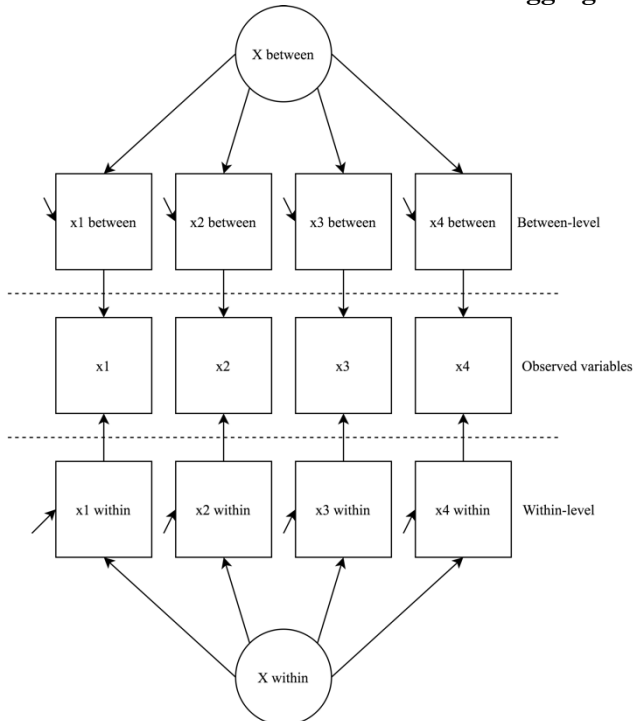## Model 1: Doubly Manifest
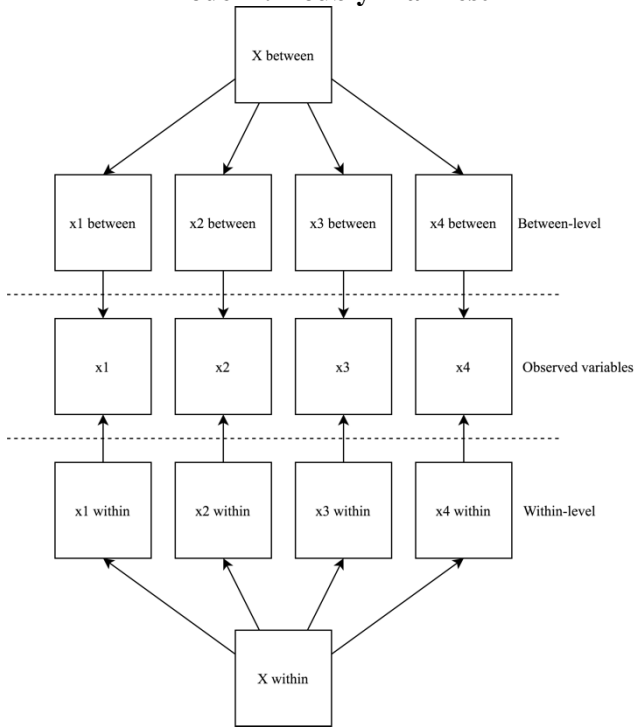
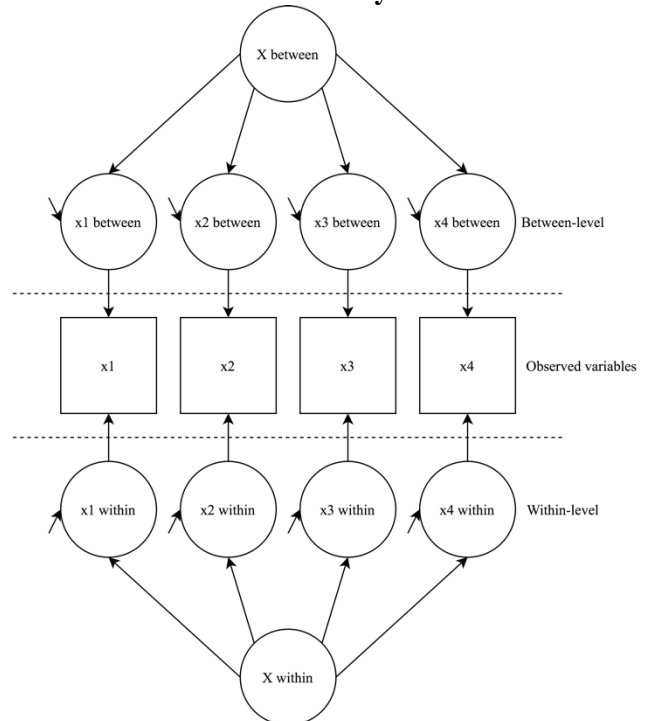## Model 2: Manifest-Measurement/Latent-Aggregation
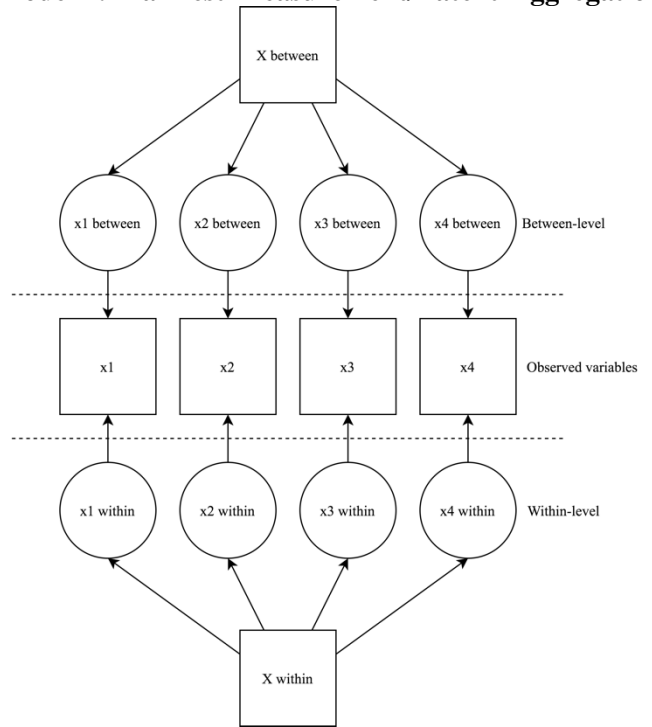
Model 3: Latent-Measurement/Manifest-Aggregation

Model 4: Doubly Latent

Figure 1. Graphic path diagrams of four MSEM measurement models. Circles represent latent variables; squares represent observed or manifest variables. Within and between levels are separated by a dashed line.
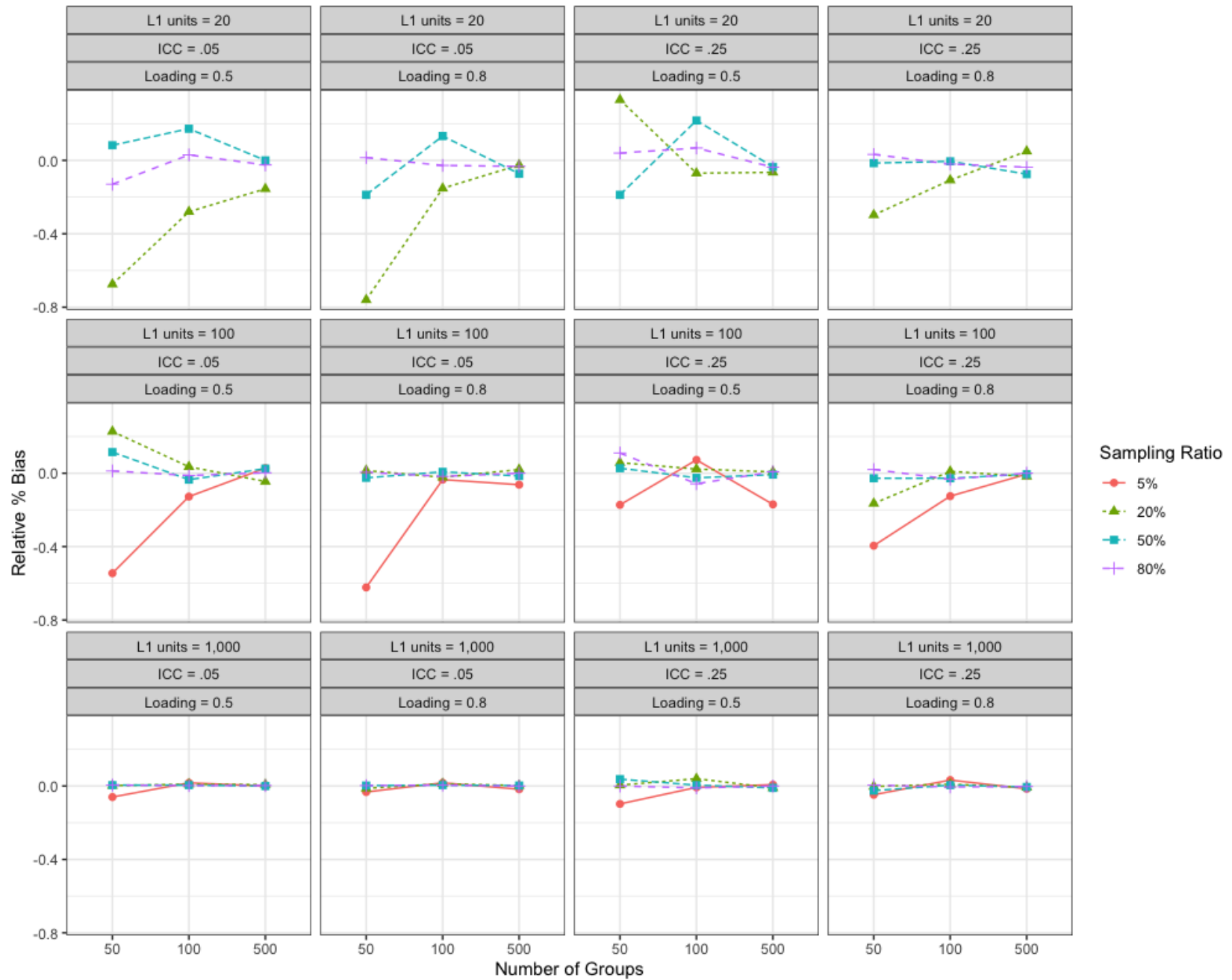
Figure 2: Relationship between the bias of estimates of the factor loadings and the sampling ratio, with varying combinations of facets.
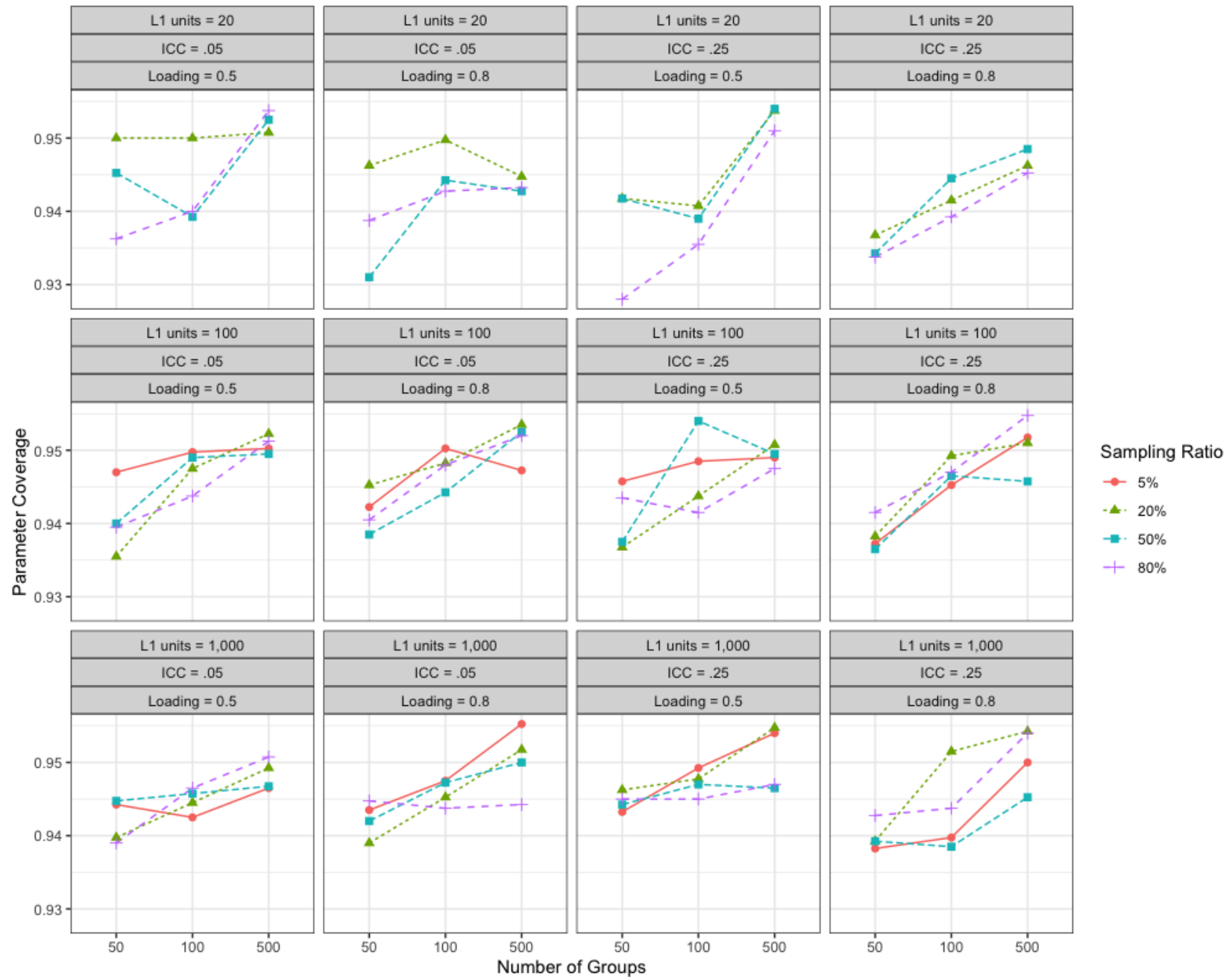
Figure 3: Relationship between the parameter coverage of estimates of the factor loadings and the sampling ratio, with varying combinations of facets.
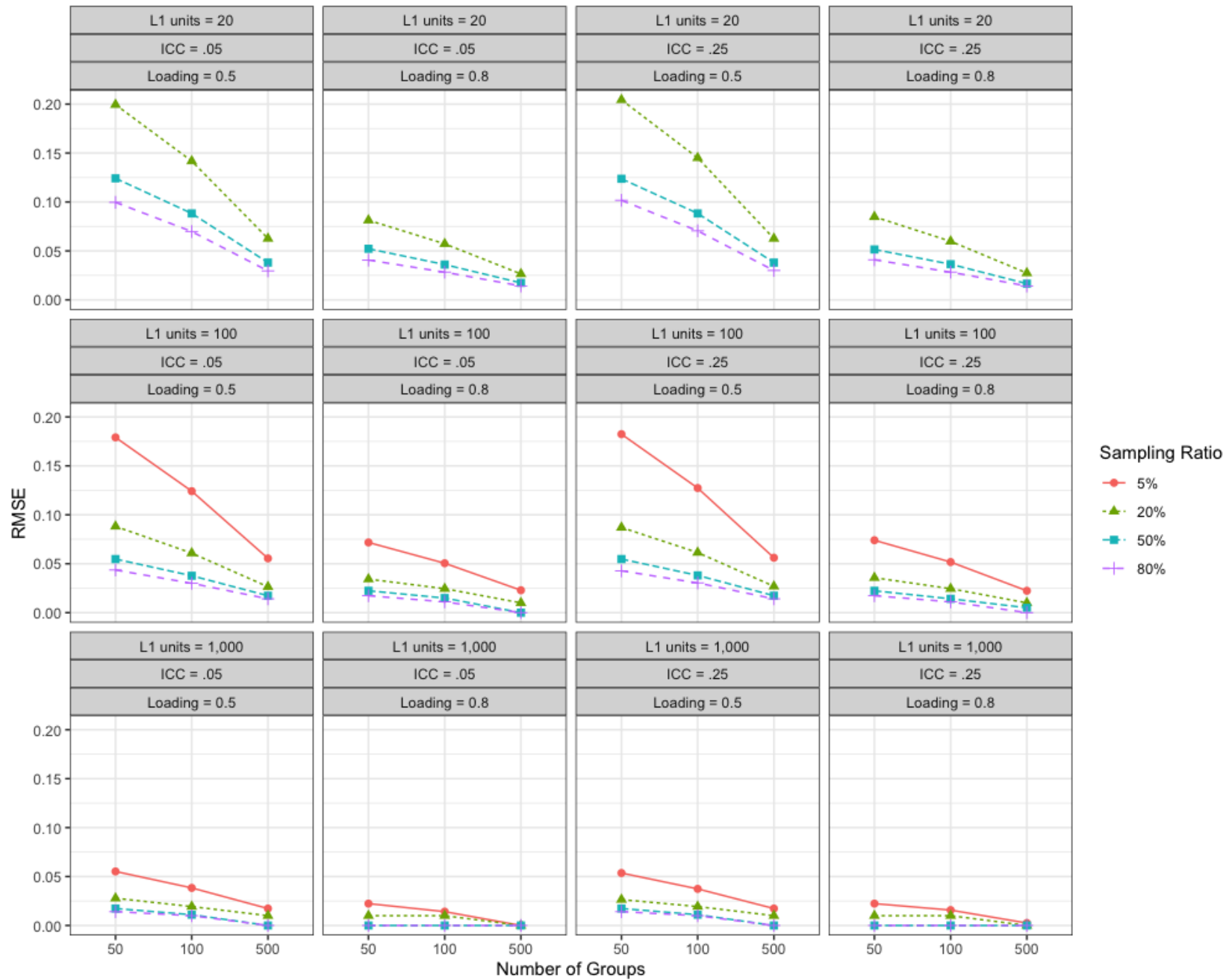
Figure 4: Relationship between the RMSE of estimates of the factor loadings and the sampling ratio, with varying combinations of facets.

Appendix A. Example Mplus Input File Used to Generate Population Data
title:
Example Mplus Input File Used to Generate
Population Data Using the Doubly Latent Model

montecarlo:
seed = 123;
names = x1-x4; !4 observed indicators at L1
ncsizes = 1;
csizes = 100(100);
nobservations = 10000;
nreps = 10; !example for 10 replications
repsave = all;
save = rep*.dat;

model population:
%within%
x1-x4*22.8; !L1 residual variance of indicators
Factor_w BY x1-x4*1; !L1 factor loadings
Factor_w*1; !L1 factor variance
[Factor_w*0]; !L1 factor mean

%between%
x1-x4*1.2; !L2 residual variance of indicators
Factor_b BY x1-x4*1; !L2 factor loadings
Factor_b*1; !L2 factor variance
[Factor_b*0]; !L2 factor mean

analysis:
type = twolevel basic;

Appendix B. Example Mplus Input File Used to Analyze Analytic Sample Datasets
title:
Example Mplus Input File Used to Analyze Analytic
Sample Datasets Using the Doubly Latent Model

data:
type = montecarlo;
file = part2_replist.dat; !name of textfile containing names of
                            !each analytic sample dataset
variable:
names = x1-x4 schid;
cluster = schid;

model:
%within%
x1-x4*22.8;
Factor_w BY x1*1(1) !Constrain factor loadings to be invariant across levels
x2*1(2)
x3*1(3)
x4*1(4);
Factor_w@1; ! L1 factor variance constrained for identification

%between%
x1-x4*1.2;
Factor_b BY x1*1(1) !Cross-level invariance
x2*1(2)
x3*1(3)
x4*1(4);

Factor_b*1; ! L2 factor variance freely estimated

analysis:
type = twolevel;
estimator = mlr;

Appendix C. Calculating Standardized Factor Loadings for Generating Data

Let variable U represent a latent variable measured by p multiple indicators. A single indicator X_{pij} for item p for person i in group j may be decomposed as (Lüdtke et al., 2011):

$$X_{pij} = \lambda_{pW} x_{Wij} + \epsilon_{ij} + \lambda_{pB} x_{Bj} + \epsilon_j; p = 1, \dots, P$$

where λ_{pW} and λ_{pB} are the within- and between-factor loadings, respectively; x_{Wij} and x_{Bj} are the unobserved true scores at L1 and L2, respectively, and ϵ_{ij} and ϵ_j are the residuals at L1 and L2, respectively. Both x_{Wij} and x_{Bj} are assumed to be normally distributed with means of zero and $\mathrm{Var}(x_{Wij}) = \sigma^2$ and $\mathrm{Var}(x_{Bj}) = \tau^2$, respectively. For generating data, unstandardized factor loadings for each indicator at both levels were set equal to 1. Similarly, measurement error variances for each item were set to σ_e^2 and τ_e^2 at L1 and L2, respectively. As a result of parallel measures at both levels, the observed variance of a single indicator may be decomposed as:

$$Var(X_{ij}) = \sigma^2 + \sigma_e^2 + \tau^2 + \tau_e^2$$

Now, the standardized factor loading for a single indicator at L1 is:

$$\lambda_{W,std} = \frac{\sigma^2}{\sigma^2 + \sigma_e^2}$$

while the standardized factor loading for a single indicator at L2 is:

$$\lambda_{B,std} = \frac{\tau^2}{\tau^2 + \tau_e^2}$$

The standardized factor loadings were set to be equal at L1 and L2. Moreover, the total variance of a single predictor variable X was assumed to equal 1, where $\mathrm{Var}(x_{Bj}) = \tau^2$ and $\mathrm{Var}(x_{Wij}) = 1 - \tau^2$. Then, the item residual variance for a single item at L1 is specified as:

$$\sigma_e^2 = \frac{(1 - \tau^2)(1 - \lambda_{W,std})}{\lambda_{W,std}}$$

while the item residual variance for a single item at L2 is:

$$\tau_e^2 = \frac{\tau^2(1 - \lambda_{B,std})}{\lambda_{B,std}}$$

For example, for simulation conditions with an intraclass coefficient (ICC$_O$) = .25 and

standardized factor loadings = 0.8, the residual variances at L1 and L2, respectively, are:

$$\sigma_e^2 = \frac{.75 * (1 - 0.8^2)}{0.8^2} = 0.42$$

and

$$\tau_e^2 = \frac{.25 * (1 - 0.8^2)}{0.8^2} = 0.14$$