The Dissertation Committee for Young Ri Lee
Certifies that this is the approved version of the following Dissertation:

A Comparison of Methods for Centering Covariates in Cross-Classified Random Effects Models

Committee:

S. Natasha Beretvas, Supervisor

James E. Pustejovsky, Co-Supervisor

Tiffany A. Whittaker

Brian T. Keller

A Comparison of Methods for Centering Covariates in Cross-Classified Random Effects Models

by

Young Ri Lee

Dissertation

Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

The University of Texas at Austin
August 2023

Acknowledgments

I would like to express my deepest gratitude to my advisors for their invaluable guidance and support throughout the duration of the writing and completion of this dissertation. Dr. Tasha Beretvas, your constant encouragement that I embrace new challenges was instrumental in my growth. Working as your research assistant provided me with invaluable learning experiences, and I was honored to have had you as my advisor and mentor during my research journey. Dr. James Pustejovsky, I am deeply grateful for the wealth of knowledge you shared. The time I spent in Pusto lab with my wonderful colleagues was enjoyable, and the advice and insights you provided during our meetings were immensely beneficial. I was indeed fortunate to have had you advise me throughout my doctoral program.

Dr. Tiffany Whittaker, I am extremely thankful for your unwavering helpfulness as my committee member. You made me feel welcomed within QM, and your constant support and invaluable advice from the qualifying exam to the dissertation empowered me. I would also like to sincerely thank committee member Dr. Brian Keller for providing valuable advice to enhance my dissertation. Your feedback and your new perspectives made my work even more robust. Thank you.

I am also grateful to my dear QM program colleagues for their support, especially during the challenging times of the COVID-19 pandemic. My lovely friends, Byun, Sam, Denti, and Seungmin, your presence and interactions guaranteed me stability and comfort. Thank you for being there for me. And my beloved family, thank you for praying for me. Your love and encouragement were my driving force throughout this academic journey. Finally, I want to express my heartfelt appreciation and love to my husband, Dongha, who has been a significant source of motivation and unwavering support throughout my journey. I would not have come this far without you.

A Comparison of Methods for Centering Covariates in Cross-Classified Random Effects Models

Young Ri Lee

The University of Texas at Austin

Supervisors: S. Natasha Beretvas and James E. Pustejovsky

The cross-classified random-effects model (CCREM) is used to handle cross-classified data in which units are nested within multiple higher-level dimensions that are not clustered within each other. The focus of interest in this study is the exogeneity assumption in CCREM, which refers to the assumed independence between covariates and random effects at level-2. If the exogeneity assumption is violated, it affects the robustness of the statistical inferences made when estimating the CCREM. Certain methods for centering a covariate can reduce the impact of violating exogeneity. For unbalanced cross-classified data, Raudenbush (2009) proposed the general model of the adaptive centering approach using cluster-mean centering. However, there are several alternatives in addition to this model, including the correlated random effects (RE) model, cell-mean centering, fixed effects (FE) using cluster robust variance estimation (CRVE), and the FE-RE hybrid model. The correlated RE model explicitly models between-cluster variability of the level-1 covariate as level-2 predictors, simultaneously estimating both within- and between-cluster effects. Another approach called cell-mean centering centers covariates around the cell mean instead of the cluster mean and considers the interaction between the two dimensions of the data. If a researcher is interested primarily in level-1 covariates, the FE approach has often been used for handling violations of exogeneity (Wooldridge, 2010). The FE model can be used along with two-way CRVE, an extension of one-way CRVE

that accounts for the dependence of errors within clusters (Cameron et al., 2011). The final alternative is an FE-RE hybrid model, which incorporates the FE and RE approaches by modeling one dimension as fixed effects and the other dimension as random effects. This approach requires fewer assumptions while benefiting from the use of the RE model for the selected dimension. However, covariate-centering strategies have only been examined for the hierarchical linear model, not for the CCREM. Thus, extended research on CCREM is needed to demonstrate and evaluate the impact of centering options on the model's performance and statistical inferences. In this dissertation, I first reviewed the current practice of centering with the CCREM and described the benefits and limitations of covariate centering methods with the CCREM. Next, I presented the results of two empirical applications comparing the use of different centering alternatives. Then, I conducted a systematic review examining how assumptions were tested and how centering was used when estimating the CCREM in applied education and social science research. Finally, I performed a simulation study to compare the performance of alternative centering approaches in scenarios in which the exogeneity assumption is violated.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the fields of educational and behavioral science, hierarchical data are a frequently encountered data structure in which lower-level units belong to higher-level clusters. In cross-sectional data, for example, students nested in schools create two-level hierarchical data. Longitudinal data through repeatedly measured scores nested within participants offer another example of hierarchical data. In hierarchical data, the level-1 observations within the same cluster correlate with each other. Educational researchers typically use a hierarchical linear model (HLM) to analyze hierarchical data to account for these dependencies between observations within a cluster (Raudenbush & Bryk, 2002).

As a generalization of hierarchical data, cross-classified data entail data in which the lower-level units are simultaneously nested within two or more clusters although the higher level clustering units are not nested within each other. For example, students are often nested within schools and neighborhoods simultaneously (e.g., Abdel Magid et al., 2021; Nuño and Katz, 2019; Pedersen et al., 2018). In this case, students who live in the same neighborhoods go to different schools although it is not the case that a unique set of neighborhoods feed into each school nor that a unique set of schools draw students from a single neighborhood. Instead, students are clustered within schools and within neighborhoods but schools are not clustered within neighborhoods nor vice-versa. Students can then be considered cross-classified by school and neighborhood. Other examples include cross-classified data structures for psychological response process data, where response times can be nested within both respondents and items simultaneously (Pae et al., 2020; Rios & Soland, 2022). In social psychology data, such as the social relations model, interpersonal perception is nested within two individuals called the target and perceiver (Schmidt et al., 2021).

If the HLM is used to analyze cross-classified data, one of the cluster dimensions of the cross-classified data is inevitably ignored because the HLM considers only one clustering dimension. As a result, use of the HLM with cross-classified data can yield underestimated standard error (SE) estimates for level-2 predictors and overestimated random effects variance components when ignoring the cross-classified nature of the data (Fielding & Goldstein, 2006; Gilbert et al., 2016; Luo & Kwok, 2009, 2012; Meyers & Beretvas, 2006; Park et al., 2017; Raudenbush & Bryk, 2002). Thus, cross-classified data can be analyzed using cross-classified random effects modeling (CCREM), an extension of the HLM (Goldstein & Sammons, 1997; Raudenbush, 1993; Van den Noortgate et al., 2003). The CCREM is designed to recognize the multiple cross-classified factor dimensions by modeling the associated random effects and can be used to include the cross-classified cluster-level covariates in the model (Raudenbush & Bryk, 2002).

The HLM and CCREM are both considered random effects (RE) models, which are premised on several strong assumptions. One of the assumptions in the RE model is that all covariates in the model are independent of the random effects, called the exogeneity assumption. If the covariates are correlated with any of the random effects, the RE models may provide inconsistent estimates (Greene, 2018).

In the HLM, cluster-mean centering has been developed to reduce the influence of violations of the exogeneity assumption. Under cluster-mean centering, the cluster-mean-centered covariate is included in the model instead of the uncentered covariate. The centered covariate no longer correlates with level-2 random effects in this case. Thus, cluster-mean centering handles violation of the exogeneity assumption (Enders & Tofighi, 2007; Kreft et al., 1995). The coefficient estimate of the cluster-mean-centered covariate represents the within-cluster effect of the covariate, which is the association between the covariate and the outcome at level-1 (Bell & Jones, 2015). In this sense, cluster-mean centering is also referred to as the *within-RE model* (Bell & Jones, 2015).

Another alternative is a *hybrid model* illustrated in Allison (2009) and Hamaker and Muthén (2020). Similar to the within-RE model, the hybrid model includes the cluster-mean-centered covariate but also includes the cluster mean as a level-2 predictor in the model. In the hybrid model, the coefficient estimate for the cluster-mean-centered covariate offers the within-cluster effect, whereas the coefficient estimate of the cluster mean predictor provides the between-cluster effect of the covariate. Also, the hybrid model captures the potential correlation between the random effects and the covariate by including the cluster mean predictor, thereby handling the violation of the exogeneity assumption. Thus, econometricians refer to the hybrid model as the *correlated RE* (CRE) model (Wooldridge, 2013, Chapter 14.3).

Using the reparameterization, the CRE model performs the same with the uncentered covariate and cluster-mean (Mundlak, 1978). Despite the lack of cluster-mean centering, the coefficient of the uncentered covariate still corresponds to the within-cluster effect of the covariate. The coefficient of the cluster mean predictor corresponds to the contextual effect, which indicates the difference between the within- and between-cluster effects (Bell & Jones, 2015; Kreft et al., 1995).

The centering alternatives described above are described in many textbooks on multilevel modeling (Hox and Roberts, 2011, Chapter 15; Hox et al., 2017, Chapter 4; Raudenbush and Bryk, 2002, Chapter 5; Snijders and Bosker, 2012, Chapter 4) and have been discussed extensively in the literature (Bell & Jones, 2015; Feaster et al., 2011; Hamaker & Muthén, 2020; Hoffman, 2019; Paccagnella, 2006; Raudenbush, 1989). However, these works almost exclusively deal with hierarchical data. Very few studies have proposed methods for handling potential endogeneity in cross-classified data. One exception is Raudenbush (2009), who proposed a general adaptive centering estimator for analyzing unbalanced cross-classified data. Like cluster-mean centering in HLM, the adaptive centering approach estimates within-cluster effects of the covariate. However, centering alternatives for cross-classified data warrant further investigation in the following two respects.

First, compared to the numerous centering alternatives for HLM (i.e., the within-RE and CRE models), the current alternatives for the CCREM only include cluster-mean centering through the general adaptive centering model (i.e., within-RE model). In other words, the researcher can only obtain the within-cluster effects of the covariates in CCREM. However, depending on the topic of the study, the between-cluster or contextual effects of the covariates in CCREM might offer researchers richer information about the association between the covariates and the outcome. A cross-classified version of the CRE model would enable in-depth exploration of covariates' between-cluster or contextual effects, but has not yet been demonstrated or evaluated.

Second, the cell interaction between clustering dimensions has not been fully considered in the context of cross-classified data. With cross-classified data, the cell refers to the combination of values for both clustering dimensions, such as the set of students who live in the same neighborhood and attend the same school. Shi et al. (2010) recommended including the random interaction effect between cross-classified factors when modeling random effects in the CCREM. In this vein, it might be reasonable to assume that covariates in cross-classified data also have a cell interaction effect between clustering dimensions in addition to the main clustering dimension effects. Therefore, I introduce the idea of *cell-mean centering*, which uses the cell mean rather than the cluster mean as the reference value for the centering and considers the interaction effects between clustering dimensions of the covariate. To our knowledge, no previous study has evaluated the benefits and limitations that cell-mean centering entails.

Other fields, such as econometrics and politics, have widely used the fixed effects (FE) model over the RE model (McNeish & Kelley, 2019; McNeish & Stapleton, 2016). When Petersen (2009) surveyed more than 200 economic studies, the studies were found to use the FE model (29%) more frequently than the RE model (less than 3%). The FE model controls the unobserved effects of clusters by including dummy coded fixed effect indicators for each cluster in the model and does not require the exogeneity assumption (Wooldridge, 2013). In balanced data and even in unbalanced data, inferences about a level-1 covariate under the FE model are equivalent to those made based on the within-RE model and the CRE model (Raudenbush, 2009; Wooldridge, 2013, Chapter 14.3). Even with cross-classified data, the FE model can account for the clusters' multiple dimensions by including cluster indicators for each cross-classified dimension, as well as interaction indicators between dimensions.

The cluster indicators in the FE model might not capture all the dependencies within clusters. In that case, cluster-robust variance estimation (CRVE) can be used to estimate the FE model, which statistically corrects the SEs of the coefficient for the covariate based on the residuals in the working model. Assuming a large number of clusters, CRVE produces asymptotically consistent SEs for the coefficients (McNeish et al., 2017; White, 1984). As a generalization of the one-way CRVE, Cameron et al. (2011) and Thompson (2011) proposed the two-way CRVE for cross-classified data. The two-way FE-CRVE eliminates the correlation between the covariates and the level-2 random effects and reduces the negative impact of endogeneity while providing consistent covariate coefficient SEs. However, the FE model does not provide random effects variance component estimates nor appropriately estimate the level-2 covariate coefficient provided in RE models.

Further, it is possible to utilize a hybrid or intermediate model, which combines both FE and RE, to analyze data with two cross-classified dimensions. This *FE-RE hybrid model* models clusters as FEs for one dimension of the cross-classified data and uses the RE model for the other dimension's clusters. This allows the estimation of the cluster-level covariate's coefficient or random effects for the RE-treated dimension while maintaining control over both dimensions. The FE-RE hybrid model requires fewer assumptions than the CCREM with an uncentered or grand-mean-centered covariate, because the hybrid model uses the RE model for only one of the two dimensions.

In this study, I explored the performance of several strategies for centering individual-level (i.e., level-1) covariates, including not centering the covariate, grand-mean centering, cluster-mean centering, cell-mean centering, use of FE-CRVE, and the FE-RE hybrid model on reducing the impact of violating the exogeneity assumption when estimating the within- and between-cluster effect of a level-1 covariate in two-level cross-classified data. I focused on unbalanced cross-sectional data widely seen in educational and behavioral science research. The study did not consider longitudinal data because longitudinal data typically offer close to balanced data, and the level-1 (e.g., time-level) covariate in longitudinal data is often centered around a certain fixed time point (c.f. Biesanz et al., 2004; Singer & Willett, 2003). Future research might consider how best to center covariates in cross-classified longitudinal data.

In the following sections, I first review and discuss alternatives for handling endogeneity in cross-classified data. Here I propose methods that extend the previously suggested centering choices for cross-classified data. These methods offer between-cluster or contextual effects of covariates and take into account the cluster-interaction effects of the covariates using cell-mean centering. I use empirical data to illustrate the types of information each approach provides and the different coefficients and SEs that they can yield. Next, I conduct a systematic review to demonstrate how the violation of the endogeneity assumption has been addressed and how centering has been used in applied CCREM analyses. Finally, I conduct a Monte Carlo simulation study to compare the performance of each method under various conditions for cross-classified data. Through this research, I intend to help applied researchers working with cross-classified data by describing and assessing a broad range of possible covariate-centering approaches for handling endogeneity.

Chapter 2

Literature Review

This section describes approaches to controlling potential endogeneity problems when analyzing cross-classified data. First, I review the CCREM and explain the endogeneity problem that can be encountered. I then describe alternative approaches for handling endogeneity, including the centering approaches in CCREM, the use of the FE model paired with CRVE, and the hybrid approach to use both FE and RE. For ease of explanation, I first explain the methods for hierarchical data and extend them to methods using cross-classified data.

2.1 Random Effects Models

2.1.1 Hierarchical Linear Model

Unconditional HLM

I begin by describing a two-level unconditional random intercept HLM that models the intercept as varying across clusters. For an illustration of a hierarchical data structure, I use an example of students clustered within schools. Suppose a hierarchical data structure where students $i \in 1, 2, ..., n_j$ are nested within schools $j \in 1, 2, ..., J$. The total number of students is $N = \sum_{j=1}^{J} n_j$. Following the notation in Raudenbush and Bryk (2002), the level-1 equation for HLM is

$$Y_{ij} = \beta_{0j} + e_{ij}, \tag{2.1}$$

where Y_{ij} is the outcome for student i in school j, β_{0j} is the average student outcome for the school j, and the error term e_{ij} is the level-1 error of student i in school j. The error terms are assumed to be independent of each other and to follow a normal distribution with a mean of zero and a constant variance of σ^2.

At level-2, the random intercept HLM models the average outcome β_{0j} to vary across schools using random effects:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \tag{2.2}$$

where γ_{00} is the overall average of the student outcome across schools and u_{0j} is the level-2 random effect independent of each other and independent from level-1 error, e_{ij}. The level-2 random effect is normally distributed with a mean of zero and a variance of τ_{00}. The combined models 2.1 and 2.2 provide the full model:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \tag{2.3}$$

Using the variance components of level-1 and level-2 errors, the proportion of the total variance in the outcome that is accounted for by variability between clusters can be calculated using the intra-class correlation coefficient (ICC):

$$\rho_{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \tag{2.4}$$

The value of the ICC captures the correlation between two randomly selected individuals from a single cluster.

HLMs are typically estimated using maximum likelihood (ML) or restricted maximum likelihood (REML) estimators. The ML estimator is a unique, minimum-variance, and unbiased estimator for fixed effects (Raudenbush & Bryk, 2002, Chapter 3). Given a large number of clusters, the ML estimator yields consistent and asymptotically efficient variance component estimates (Goldstein, 1986; Longford, 1987; Raudenbush & Bryk, 2002). However, with a small number of clusters, REML produces more accurate random effects variance component estimates by considering the sampling variation of the fixed effects estimator (Goldstein, 2011; Mason et al., 1983; Raudenbush & Bryk, 1986; Snijders & Bosker, 2012, Chapter 4.7). Specifically, REML incorporates the loss in the degree of freedom due to estimating fixed effects estimates for the covariates and provides an adjusted variance component estimate. Thus, when the number of covariates is small, the difference between ML and REML is small, but the difference between ML and REML increases as the number of covariates increases (Snijders and Bosker, 2012, Chapter 4.7; Raudenbush and Bryk, 2002, Chapter 3).

Conditional HLM

Based on the research questions, the conditional HLM includes covariates at the corresponding level of the model. For example, in the previous hierarchical data structure where students are nested within schools, the level-1 covariates can be student characteristics, such as gender or age. The level-2 covariates might be cluster characteristics, such as school resources or school type. Considering the random intercept model of HLM with a level-1 covariate X_{ij}, the level-1 equation of the conditional model is

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, \tag{2.5}$$

where β_{0j} is the conditional intercept for school j, β_{1j} is the slope coefficient for the covariate X_{ij}, and e_{ij} is a level-1 error. The level-1 errors are also assumed mutually independent and normally distributed with a mean of zero and a homogeneous conditional variance of σ^2.

In a conditional HLM, both the conditional intercept β_{0j} and slope β_{1j} for the covariate can be modeled as varying across level-2 clusters. As the simplest form, however, I assume only the intercept β_{0j} to be varying across schools. The remaining slope β_{1j} is assumed to be fixed across schools. When the level-2 covariate W_j is modeled as a predictor of the schools' intercept, the level-2 model is

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j}, \\ \beta_{1j} &= \gamma_{10}, \end{aligned} \tag{2.6}$$

where γ_{00} is the overall mean of the outcomes across the clusters, γ_{01} is the regression coefficient for the level-2 covariate W_j, and u_{0j} indicates the level-2 random effect, assumed to follow a normal distribution with a mean of zero and a conditional homogeneous variance of τ_{00}. The level-2 random effects are also assumed to be mutually independent and independent of the level-1 errors. Further, the level-2 random effects and the level-1 error are not correlated with any of the newly included covariates, X_{ij} or W_j (i.e., the exogeneity assumption). In this model, the slope coefficient γ_{10} for the covariate X_{ij} is held constant across schools. The combined equation for the conditional model becomes

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + u_{0j} + e_{ij}. \tag{2.7}$$

For the conditional model, the conditional ICC is calculated using Equation 2.4 where σ^2 and τ_{00} now provide the respective variances of conditional level-1 errors and conditional level-2 random effects.

In educational research, the presence of clustered data, which includes schools, teachers, and districts, is quite common. Consequently, it becomes crucial to address the inherent dependencies within such data. HLM offers one approach to handle these dependencies. Additionally, HLM allows researchers to estimate the impact of level-2 covariates on the outcome variable and explore random effects at the cluster level, representing unexplained variance between clusters.

2.1.2 Cross-Classified Random Effects Model

Unconditional CCREM

CCREM extends HLM to model randomly varying intercepts for each of two or more cross-classified clustering factors. I use two clustering dimensions in the illustrations (e.g., schools and neighborhoods) and review the two-level unconditional random intercept CCREM before including any covariates in the model. Given a cross-classified data structure where individuals $i \in 1, 2, ..., n_{jk}$ within cell jk are cross-classified by schools $j \in 1, 2, ..., J$ and neighborhood $k \in 1, 2, ..., K$, the level-1 equation of the CCREM is

$$Y_{i(jk)} = \beta_{0(jk)} + e_{i(jk)}, \tag{2.8}$$

where $Y_{i(jk)}$ indicates the outcome variable for student i within school j and neighborhood k, $\beta_{0(jk)}$ indicates the average outcome for those who attending school j and living in neighborhood k. The last term $e_{i(jk)}$ denotes the level-1 errors, which are assumed to be normally distributed with a mean of zero and constant variance of σ^2. The indices for two cross-classified clustering dimensions (e.g., jk) are shown as subscripts in parentheses. The shared parentheses are used to symbolically represent that neither cluster is nested within the other.

The CCREM allows the intercept $\beta_{0(jk)}$ to vary across the two cross-classified clustering dimensions. The level-2 equation of the unconditional CCREM is

$$\beta_{0(jk)} = \gamma_{000} + b_{0j0} + c_{00k} + d_{0jk}, \tag{2.9}$$

where γ_{000} is the overall average of the outcome variable across all schools and neighborhoods. The terms b_{0j0} and c_{00k} indicate the random effects for each clustering dimension (e.g., schools and neighborhoods). These random effects are assumed to be normally distributed with a mean of zeros and variances of τ_{j00} and τ_{k00}. The last random effect $d_{0jk} \sim N(0, \tau_{(jk)00})$ is the random interaction effect between the two cross-classified clustering dimensions, indicating the deviation from the cell mean calculated by the two cluster means from each dimension and the grand mean (Raudenbush & Bryk, 2002). If this random interaction effect is omitted, the level-2 variance components of the random effects can be inflated, whereas the coefficient estimators for the predictors remain unbiased (Shi et al., 2010).

Combining the level-1 and level-2 equations, the full unconditional CCREM is

$$Y_{i(jk)} = \gamma_{000} + b_{0j0} + c_{00k} + d_{0jk} + e_{i(jk)}. \tag{2.10}$$

As in HLM, the level-2 errors b_{0j0}, c_{00k}, and d_{0jk} and the level-1 error $e_{i(jk)}$ are assumed to be mutually independent. Instead of the HLM's ICC, there is an intra-unit correlation coefficient (IUCC) for CCREM that captures the degree of clustering. For example, the intra-school correlation coefficient is estimated using:

$$\rho_{IUCC_j} = \frac{\tau_{j00}}{\tau_{j00} + \tau_{k00} + \tau_{(jk)00} + \sigma^2}, \tag{2.11}$$

indicating the correlation between outcome variables from two students within the same school but who live in different neighborhoods. In the case of calculating an intra-neighborhood or intra-interaction correlation coefficient, the numerator in Equation 2.11 is τ_{k00} or $\tau_{(jk)00}$ instead of τ_{j00}.

Conditional CCREM

In the conditional CCREM, level-1 and level-2 covariates are included. Using the same notation from the unconditional CCREM, the level-1 CCREM, including the level-1 covariate $X_{i(jk)}$, is:

$$Y_{i(jk)} = \beta_{0(jk)} + \beta_{1(jk)} X_{i(jk)} + e_{i(jk)}, \tag{2.12}$$

where $\beta_{0(jk)}$ is the conditional average of the outcome for students in school j and neighborhood k, and $\beta_{1(jk)}$ is the slope for $X_{i(jk)}$, the covariate values for a student i nested within school j and neighborhood k. The level-1 errors $e_{i(jk)}$ are assumed to follow a normal distribution, with a mean of zero and constant variance of σ^2, and are mutually independent.

The level-2 CCREM allows modeling of school and neighborhood covariates

(W_j and Z_k, respectively) as explaining variability in the intercept and the slopes:

$$\begin{aligned}\beta_{0(jk)} &= \gamma_{000} + \gamma_{010}W_j + \gamma_{002}Z_k + b_{0j0} + c_{00k} + d_{0jk}, \\ \beta_{1(jk)} &= \gamma_{100},\end{aligned} \tag{2.13}$$

where γ_{000} is the conditional intercept across all schools and neighborhoods controlling for the covariates included, γ_{010} is the coefficient for the school covariate W_j, and γ_{002} is the coefficient for the neighborhood covariate Z_k. The level-2 random effects, b_{0j0} for schools, c_{00k} for neighborhoods, and d_{0jk} for the interaction between schools and neighborhoods, are mutually independent and assumed to be normally distributed with a mean of zeros and conditional variance τ_{j00}, τ_{k00}, and $\tau_{(jk)00}$, respectively. I modeled the slope $\beta_{1(jk)}$ to be constant across the level-2 units for simplicity. In other words, the γ_{100} represents the fixed slope for the level-1 covariate $X_{i(jk)}$.

Combining level-1 and level-2, the full conditional CCREM is

$$Y_{i(jk)} = \gamma_{000} + \gamma_{100}X_{i(jk)} + \gamma_{010}W_j + \gamma_{002}Z_k + b_{0j0} + c_{00k} + d_{0jk} + e_{i(jk)}. \tag{2.14}$$

The level-1 errors and the level-2 random effects are uncorrelated with any of the included covariates (i.e., the exogeneity assumption). The IUCC is calculated by the level-1 and level-2 random effects using Equation 2.11, indicating the conditional IUCC.

The CCREM provides valuable insight into each clustering dimension that researchers might need to recognize in their data. Cluster-level covariates describing any clustering dimension can be included in the model. Covariates pertaining to school characteristics or neighborhood characteristics can be included in a single model, even though the covariates describe different clustering dimensions. Moreover, CCREM is beneficial in estimating the distribution of random effects for each clustering dimension. The variance of these random effects represents the variability between each clustering dimension (Raudenbush & Bryk, 2002, Chapter 12). In other words, the variance of these random effects describes the variation of the true clustering means for each clustering dimension around the grand mean. Finally, CCREM enables flexible modeling and provides more efficient estimates when all assumptions are met compared to the other methods, such as the FE model (Wooldridge, 2003, Chapter 14.3).

2.2 Endogeneity Problem

Despite their potential for providing rich descriptions of data, these RE models require more strict assumptions than the FE model. In particular, for the two-level HLM, the exogeneity assumption requires that the level-1 error and level-2 random effects are uncorrelated with any of the covariates. In this work, I focus only on endogeneity bias caused by the correlation between the level-1 covariate and the level-2 random effects because the independence of the level-1 covariates and error is also required with the FE model. The level-1 error endogeneity is beyond the scope of this study.

The exogeneity assumption can be violated primarily when the covariate is correlated with the level-2 random effects due to omitted variables or unobserved heterogeneity (Antonakis et al., 2021; Bell & Jones, 2015; Raudenbush & Bryk, 2002). Consider an example involving a hierarchical data structure, where students are nested within schools. When the outcome variable is student achievement, and the predictors are the quality of tutoring students receive (level-1) and school resources (level-2), the HLM equation is:

$$Achievement_{ij} = \gamma_{00} + \gamma_{10} Tutoring_{ij} + \gamma_{01} SchoolResource_j + u_{0j} + e_{ij}, \qquad (2.15)$$

where $Achievement_{ij}$ is the outcome score of student i in school j, $Tutoring_{ij}$ is the tutoring quality that student i at school j receives, and $SchoolResource_j$ indicates the resource in school j. When schools with higher resources may encourage students to receive better tutoring by connecting them with higher-quality teachers, the school resources can correlate with tutoring quality. In other words, school resources influence students' achievement and tutoring quality simultaneously. In this instance, if researchers fail to measure school resources and do not include them in HLM as a level-2 predictor, the unexplained school variance due to the school resources would be absorbed into the level-2 random effects:

$$Achievement_{ij} = \gamma_{00} + \gamma_{10} Tutoring_{ij} + \upsilon_{0j} + e_{ij}, \tag{2.16}$$

where υ_{0j} is a new random effect at level-2 that absorbs the effect of the school resource variable. Thus, the exogeneity assumption is violated because the level-2 random effect υ_{0j} is correlated with the level-1 covariate, the tutoring quality.

The exogeneity assumption in HLM has often been tested through the Hausman test, which was widely used in other fields, such as econometrics (Hausman, 1978). The Hausman test examines whether the efficient estimator from RE and consistent estimator from FE are estimating the same parameter value. If the null hypothesis is rejected, the RE and FE estimators have different biases (i.e., they are not estimating the same quantity), and the consistent FE estimator is recommended. On the other hand, when the null hypothesis is failed to reject, it suggests that the efficient RE estimator is very close to the consistent FE estimator.

However, the Hausman test has limitations when applied to test the exogeneity assumption of the RE model because comparing the RE and FE estimators does not directly evaluate the exogeneity assumption. It only tests the null hypothesis that RE and FE estimators are sufficiently close (Wooldridge, 2013, Chapter 14.2). The Hausman test also strictly assumes normality and homoscedasticity of errors in the two compared models, but these assumptions can be addressed using a robust Hausman test (Hausman, 1978; Wooldridge, 2010, Chapter 10.7.3).

If the exogeneity assumption is not met, the accuracy of estimates in the HLM is expected to be reduced (McNeish & Kelley, 2019; Petersen, 2009). Parameter estimates of the predictor's coefficients (fixed effects) are no longer statistically consistent and will be biased (Antonakis et al., 2021; Greene, 2018, Chapter 11.8; Kennedy, 2008, Chapter 18.3; Wooldridge, 2010, Chapter 9.4, 10.2). The SE of the coefficients also can be biased (LeBeau, 2013; McNeish et al., 2017; Schielzeth et al., 2020). Darandari (2004) and Maeda (2007) have demonstrated that mis-specification of an HLM due to omitted covariates yields biased parameter and random effects variance component estimates.

Compared to a two-level HLM, a two-level CCREM includes more clustering dimensions, requiring exogeneity assumptions on the random effects for both cross-classified clustering dimensions and their interaction (b_{0j0}, c_{00k} and d_{0jk}), respectively. Consider a cross-classified data example where students are nested within schools and neighborhoods simultaneously. Suppose the quality of tutoring students receive (level-1) is a function of school resources and neighborhood-level educational climate in this cross-classified data (both at level-2), CCREM equation is:

$$\begin{aligned} Achievement_{i(jk)} = \ & \gamma_{000} + \gamma_{100} Tutoring_{i(jk)} + \gamma_{010} Resource_j + \gamma_{002} Climate_k \\ & + b_{0j0} + c_{00k} + d_{0jk} + e_{i(jk)}, \end{aligned} \tag{2.17}$$

where $Achievement_{i(jk)}$ is the outcome score of student i in school j and neighborhood k, $Tutoring_{i(jk)}$ is tutoring quality that student i receives in school j and neighborhood k, $Resource_j$ is the school resource in school j, and $Climate_k$ is the education climate in neighborhood k.

If the school resource or neighborhood educational climate variable is omitted from the CCREM, the variability of the corresponding cluster-level variable explaining the tutoring quality is absorbed into the random effects. For example, if the neighborhood educational climate is omitted, the variability of the neighborhood's educational climate explaining the tutoring quality could partly be absorbed in the neighborhood random effects:

$$\begin{aligned} Achievement_{i(jk)} = \ & \gamma_{000} + \gamma_{100} Tutoring_{i(jk)} + \gamma_{010} Resource_{j} \\ & + b_{0j0} + \nu_{00k} + d_{0jk} + e_{i(jk)}, \end{aligned} \tag{2.18}$$

where ν_{00k} is a new level-2 random effect that absorbs the effect of the neighborhood education climate variable. Thus, the related random effects become correlated with the level-1 covariate, violating the exogeneity assumption. However, given that CCREM has multiple dimensions, there is no guarantee that the variability in the omitted neighborhood-level covariates is incorporated into the neighborhood random effects alone. In the previous example, I intended to demonstrate conceptually how the exogeneity assumption can be violated.

Considering violations of the exogeneity assumption in HLM affect the statistical inference of the results, it is plausible that violations of the exogeneity assumption in CCREM could have a comparable effect on the estimation results. For example, parameter estimates of the covariates' coefficients are no longer consistent and may be biased. The standard error of the parameter estimate may also be biased (LeBeau, 2013; McNeish et al., 2017; Schielzeth et al., 2020).

Several alternatives have been proposed for handling the risk of endogeneity in the RE model. A primary alternative discussed for the HLM involves centering covariates. For the HLM, centering is classified into grand-mean centering and cluster-mean centering, but only cluster-mean centering reduces the impact of endogeneity.

Another alternative for handling endogeneity in the hierarchical data involves use of the FE model, which does not require the exogeneity assumption for the relationship between random effects and covariates. The FE model is commonly used in economics and political science research, where the estimation of level-1 covariate effects is of primary focus over the effects of cluster-level covariates or of random effects. In addition, CRVE can be used to estimate the FE model to control for within-cluster dependencies that have not been handled.

The following sections describe alternatives that can reduce endogeneity issues in hierarchical data, including several types of centering methods in HLM that offer different decomposition of covariate effects and one-way FE CRVE with CRVE. Then, I discuss results from previous systematic reviews examining how endogeneity problems were handled in previous applied HLM research.

2.3 Alternatives in Hierarchical Data

This section reviews and compares a two-level HLM with uncentered covariates, a model with grand-mean centering, and cluster-mean centering. In particular, I describe the within-RE model and CRE model that uses cluster-mean centering as alternatives for handling endogeneity. Note that throughout, I am focused on two-level models.

Use of uncentered covariates means that the original values for the continuous or categorical covariates are included in the model. For example, in a two-level conditional HLM,

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + u_{0j} + e_{ij}, \tag{2.19}$$

the level-1 covariate X_{ij} and level-2 covariate W_j are uncentered covariates that were not transformed before being included as covariates in the model. In a model with uncentered covariates, the intercept is interpreted as the predicted value for a case with truly zero values on the covariates. In other words, the intercept γ_{00} in which achievement is the outcome variable would be the grand-mean achievement score for all students across schools with zeroes on both X_{ij} and W_j.

Suppose X_{ij} is the student's age and is the only covariate in the model. The intercept γ_{00} then represents the expected student achievement score for a student at zero years old, which is clearly not a meaningful parameter. In such a case, the uncentered level-1 covariate is not appropriate, and centering methods are recommended to provide a meaningful intercept.

The uncentered level-1 covariate is also problematic in that it can cause a violation of the exogeneity assumption. In hierarchical data structures, the slope estimate indicates the pooled effects of within-cluster effect $\hat{\gamma}_w$ and between-cluster effect $\hat{\gamma}_b$ of X_{ij} on the outcome. Here, the within-cluster effect refers to how one-unit differences in the level-1 (here, student) covariate are associated with a predicted difference in the outcome within that cluster and between-cluster effect refers to how one-unit differences in the mean predictor between the clusters are associated with the difference in mean outcome at the cluster level (Raudenbush & Bryk, 2002, Chapter 6).

In a typical HLM with uncentered covariates, γ estimators are generalized least squares estimators (Raudenbush & Bryk, 2002, Chapter 9), and the slope parameter assumes equal within- and between-cluster effects (Bartels, 2008; Snijders & Bosker, 2012). If the within- and between-cluster effects are truly equal, the slope estimate becomes the most efficient and unbiased estimate (Raudenbush & Bryk, 2002, Chapter 5). The slope estimate is calculated using

$$\hat{\gamma}_{10} = \hat{\gamma}_{pooled} = (1 - \eta^2)\hat{\gamma}_w + \eta^2\hat{\gamma}_b, \tag{2.20}$$

where η^2 is the proportion of variance in X_{ij} explained by the difference between clusters (Kreft et al., 1995). Consequently, the slope estimate will be the same as within- or between-cluster effects ($\hat{\gamma}_{10} = \hat{\gamma}_w = \hat{\gamma}_b$), given that the within- and between-cluster effects are equal.

However, previous research has indicated that there may be differences between the within- and between-cluster effect. For example, Palta and Seplaki (2002) used hierarchical data from the health panel survey to show that the within- and between-cluster effects were different for the effects of age and self-reported health on hospitalization. If the within- and between-cluster effects of the covariate are not equal, the slope estimates for the corresponding covariate will not fully account for either of them. The remaining variance is absorbed into the error terms, which becomes correlated with the covariate and violates the exogeneity assumption (Bell & Jones, 2015; Palta & Seplaki, 2002). In that case, the slope parameter estimates are no longer efficient and become an uninterpretable weighted average of within- and between-cluster effects (Bell & Jones, 2015; Raudenbush & Bryk, 2002).

The variance not accounted for in estimating the coefficient uniquely for the covariate at two levels produces a bias in the variance of level-2 random effects and level-1 errors (Grilli & Rampichini, 2011). Therefore, using appropriate centering in HLM is essential for reducing the influence of endogeneity and for obtaining consistent parameter estimates, including both fixed effects and the variance of the random effects (Enders & Tofighi, 2007).

2.3.1 Grand-Mean Centering

Grand-mean centering refers to including a covariate in a model after transforming its values to entail the deviation from the individual's raw score from the mean on the covariate. The covariate value's transformation can vary depending on the research questions. Researchers may use a median or other specific value as the centering constant instead of the grand mean (Hoffman, 2019; Lin et al., 2016). If the covariates are dummy variables such as gender, the total proportion of women can be the centering constant (Raudenbush & Bryk, 2002). And for longitudinal models, researchers commonly center a time variable around a certain point in time. However, in this study, I focus only on the general case where the continuous covariates are centered around the sample's grand mean.

If the level-1 covariate X_{ij} and level-2 covariate W_j are both grand-mean-centered, it can be represented as

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}) + \gamma_{01}(W_j - \bar{W}) + u_{0j} + e_{ij}, \tag{2.21}$$

where $\bar{X}$ indicates the overall means of the level-1 covariate X_{ij} and $\bar{W}$ is the overall mean of the cluster-level covariate W_j across the clusters. The intercept γ_{00} is then the expected outcome for those who have X_{ij} equal to the grand mean $\bar{X}$ and W_j equal to the grand mean $\bar{W}$ (Raudenbush & Bryk, 2002, Chapter 2). As such, grand-mean centering can offer a more meaningful interpretation of the model's intercept than the model with an uncentered covariate (Hox et al., 2017, Chapter 4). For example, if X_{ij} is the student's age and W_j is a measure of school resources, the intercept γ_{00} indicates the expected achievement scores for a student in the average age within a school with average school resources.

Grand-mean centering offers a *reparameterization* of the corresponding model with an uncentered covariate (Kreft et al., 1995; Raudenbush & Bryk, 2002, Chapter 2). Comparing a model with uncentered covariates and grand-mean centering reveals that the new intercept calculated under grand-mean centering (e.g., $\gamma_{00,grand}$) is the intercept of the uncentered model (e.g., $\gamma_{00,uncentered}$) adjusted by $\bar{X}$ and $\bar{W}$:

$$\gamma_{00,grand} = \gamma_{00,uncentered} + \gamma_{10}\bar{X} + \gamma_{01}\bar{W}. \tag{2.22}$$

On the other hand, the slope coefficients γ_{10} and γ_{01} are equivalent to the slopes in the uncentered model, and the within- and between-cluster effects of the covariates remain confounded. The variance of the random effects u_{0j}, τ_{00}, remains unchanged when the grand-mean centering is used in the model with a fixed slope. However, if the random slope is employed, the variance of the random effects may be different (Enders & Tofighi, 2007).

In short, grand-mean centering helps the interpretation of the intercept coefficient to be meaningful. However, grand-mean centering does not solve the endogeneity problem. Similar to the uncentered model, the coefficient estimates in grand-mean-centered models may be biased if covariates and cluster-level random effects are correlated.

2.3.2 Cluster-Mean Centering

Cluster-mean centering (i.e., adaptive centering) is an alternative that overcomes the endogeneity problem for the HLM. Cluster-mean centering involves the use of covariates that are centered around the cluster mean $\bar{X}_j$ to which each individual belongs. I describe two cluster-mean centering approaches for estimating different covariate effects: namely, the within-RE and CRE models.

In illustrating these models, I focus on the level-1 covariate and omit an additional level-2 covariate W_j. The covariates at the highest level (e.g., level-2) can only be treated as uncentered or grand-mean-centered and provide the same coefficient estimates regardless of the centering type, as long as cluster-mean centering is used for level-1 covariates.

Within-Random Effects Model

The within-RE model estimates the within-cluster effects of the covariate using the cluster-mean-centered covariate. For an HLM with a level-1 covariate, the within-RE model is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_j) + u_{0j} + e_{ij}, \tag{2.23}$$

where $(X_{ij} - \bar{X}_j)$ indicates the cluster-mean-centered covariate that subtracts the relevant cluster mean $\bar{X}_j$ from the individual's value on the covariate X_{ij}. By centering around the cluster mean, the within-RE model eliminates between-cluster variation from the total variation in the covariate. The within-cluster variation remains in the covariate, and the slope estimate, $\hat{\gamma}_{10}$, becomes the within-cluster effect. Suppose the level-1 covariate is the student's reading ability, and the outcome variable is the student's academic achievement. Then, the within-cluster effect of the student's reading ability is interpreted as the predicted difference in academic achievement for two students in the same school whose reading ability differs by one.

Because cluster-mean centering modifies the mean and correlation structure of covariates, other parameter estimates differ from their estimates when using the corresponding uncentered or grand-mean-centered models (Enders & Tofighi, 2007). The intercept γ_{00} is the expected outcome Y_{ij} while controlling for the cluster-mean-centered level-1 covariate $(X_{ij} - \bar{X}_j)$. Because the covariate is scaled differently, the intercept in the cluster-mean centering is not equivalent to those in the uncentered model or grand-mean centering. Also, the intercept random effects' variance components under cluster-mean centering are different from the uncentered or grand-mean centering model because the scale of the intercept becomes different (Kreft et al., 1995).

The within-RE model has the benefit of estimating the within-cluster effect while removing the correlation between the covariates and the random effects (Hofmann & Gavin, 1998). However, the drawback of this model is that it does not reveal the impact of the between-cluster variability of the covariate. If the researchers are interested in capturing the between-cluster or contextual effects of the covariate, the within-RE model might not be appropriate. Considering that the covariate in hierarchical data can sometimes include both within- and between-cluster effects, a model incorporating both effects would offer an attractive alternative.

Correlated Random Effects Model

The CRE model is another alternative that allows the covariates and random effects to be correlated (Allison, 2009; Bell & Jones, 2015; Enders & Tofighi, 2007; Hamaker & Muthén, 2020; Wooldridge, 2013, Chapter 14.3). The CRE model simultaneously provides the within- and between-cluster effects because it is derived by combining the equations for the within- and the between-cluster relationship of the covariate and the outcome variables in HLM (Raudenbush & Bryk, 2002, Chapter 5). In the two-level HLM, the within-cluster relationship of the covariate X_{ij} and the outcome Y_{ij} for a student i in school j is

$$Y_{ij} - \bar{Y}_j = \gamma_w(X_{ij} - \bar{X}_j) + e_{ij}, \tag{2.24}$$

where γ_w is the within-cluster effect of the covariate. Likewise, the between-cluster relationship of the covariate X_{ij} and the outcome Y_{ij} is

$$\bar{Y}_j = \gamma_{00} + \gamma_b \bar{X}_j + u_{0j}, \tag{2.25}$$

where γ_b indicates the between-cluster effect of the covariate. Combining Equations 2.24 and 2.25 then yields the CRE model,

$$\begin{aligned} Y_{ij} &= (\gamma_{00} + \gamma_b \bar{X}_j + u_{0j}) + \gamma_w (X_{ij} - \bar{X}_j) + e_{ij}, \\ Y_{ij} &= \gamma_{00} + \gamma_w (X_{ij} - \bar{X}_j) + \gamma_b \bar{X}_j + u_{0j} + e_{ij}, \end{aligned} \tag{2.26}$$

which provides both the within-cluster effect, γ_w, and between-cluster effects, γ_b, of the covariate, X_{ij}. Using the example as in the within-cluster effect above, the between-cluster effect of the covariate (i.e., student's reading ability) indicates the predicted difference in a school's average academic achievement associated with the one-unit difference in the school's average reading.

Another way of expressing the CRE model is called the Mundlak model:

$$Y_{ij} = \gamma_{00} + \gamma_w X_{ij} + \gamma_c \bar{X}_j + u_{0j} + e_{ij}, \tag{2.27}$$

where X_{ij} is the uncentered covariate and γ_c is the contextual effect that captures the difference between the within- and between-cluster effects ($\gamma_c = \gamma_b - \gamma_w$; Bell and Jones, 2015; Kreft et al., 1995; Mundlak, 1978). This contextual effect is also referred to as the compositional effect (Raudenbush & Bryk, 2002) or the incremental between-person effect (Hoffman, 2019). Considering the previous example, the contextual effect represents the predicted difference in the academic achievement outcome between two students who have the same reading ability score but are nested within schools where their schools' average reading ability score differs by one unit (Raudenbush & Bryk, 2002, Chapter 5).

Rearranging Equation 2.27 proves that it is identical to Equation 2.26 when the slope parameter is modeled as fixed at cluster-level (Kreft et al., 1995):

$$\begin{aligned}
Y_{ij} &= \gamma_{00} + \gamma_w X_{ij} + (\gamma_b \bar{X}_j - \gamma_w \bar{X}_j) + u_{0j} + e_{ij}, \\
Y_{ij} &= \gamma_{00} + (\gamma_w X_{ij} - \gamma_w \bar{X}_j) + \gamma_b \bar{X}_j + u_{0j} + e_{ij}, \\
Y_{ij} &= \gamma_{00} + \gamma_w (X_{ij} - \bar{X}_j) + \gamma_b \bar{X}_j + u_{0j} + e_{ij}.
\end{aligned} \tag{2.28}$$

However, when modeling a random slope for the level-1 covariate, the random slope variance of the two models is different because the level-1 predictors to which the random slope is applied are different (Hoffman, 2019; Kreft et al., 1995).

Compared to the within-RE model, the CRE model directly estimates the between-cluster or contextual effects of a covariate on the outcome. If a researcher wishes to examine how the effects of a covariate affect the outcome variable differs at the individual and cluster levels, the ability to estimate the between-cluster or contextual effect would be essential. Also, a test of statistical significance can be performed for the between-cluster or contextual effects in the CRE model. Specifically, the testing for the contextual effects functions exactly the same as the Hausman test (Snijders & Bosker, 2012, Chapter 4.6). Moreover, the CRE model provides different random effects variance components compared to the within-RE model, as the cluster means serve as additional control variables (Kreft et al., 1995).

Overall, cluster-mean centering approaches, including the within-RE model and CRE model, have several important benefits (Enders & Tofighi, 2007). First, cluster-mean centering removes the dependence between cluster random effects and the covariate. In other words, cluster-mean centering helps address endogeneity with respect to the relevant covariate and offers the added benefit of estimating the coefficients for level-2 covariates and the random effects variance component. However, since the effect of centering is only for the centered variables, this advantage assumes there is no endogeneity problem for the other variables.

Also, within- and between-cluster effects need not be assumed to be the same when using cluster-mean centering because these effects are estimated separately. Not centering or grand-mean centering provides an efficient estimator when the two effects are the same in actual data. However, when the two effects are different, cluster-mean centering provides a less biased within-cluster effect than a model with an uncentered or grand-mean-centered covariate (Raudenbush & Bryk, 2002, Chapter 5).

2.3.3 One-Way FE-CRVE

As an alternative approach to centering, the FE model offers another option for analyzing hierarchical data (Allison, 2005, 2009). The FE model provides consistent and unbiased parameter estimates for the within-cluster effect while avoiding the exogeneity assumption. For example, in the hierarchical data where students $i \in 1, 2, ..., n_j$ are nested within schools $j \in 1, 2, ..., J$, the FE model with a covariate is

$$Y_{ij} = \gamma_{10} X_{ij} + \sum_{p=1}^{J} \gamma_{0p} D_p + e_{ij}, \tag{2.29}$$

where γ_{10} is the regression coefficient for the covariate X_{ij}, and γ_{0p} is a cluster-level intercept for the dummy variable D_p representing each cluster indicator. The term e_{ij} is a random error for an individual i in a cluster j. The FE model can be estimated using OLS estimation.

Note that γ_{0p} estimated in the FE model is fixed effects, not random effects, as in HLM. In other words, the FE model assumes clusters are not randomly sampled from the population and models cluster effects as fixed effects. In this vein, the FE model has often been used by applied researchers who want to estimate the impact of specific clusters, such as countries or companies.

However, if the number of clusters is large, it might be infeasible to estimate the corresponding coefficients for all cluster-specific indicators. To avoid such unnecessary calculations, the FE model uses *fixed effects estimator* or *within estimator*, which is a pooled OLS that removes cluster-level variability and avoids the estimation of all cluster indicators (Greene, 2018, Chapter 11.4; Wooldridge, 2013, Chapter 14.1).

Using the FE estimator, all the cluster-level variability is removed from the FE model:

$$\ddot{Y}_{ij} = \gamma_{10}\ddot{X}_{ij} + \ddot{e}_{ij}, \tag{2.30}$$

where $\ddot{Y}_{ij} = Y_{ij} - \bar{Y}_j$, $\ddot{X}_{ij} = X_{ij} - \bar{X}_j$, and $\ddot{e}_{ij} = e_{ij} - \bar{e}_j$. If Equation 2.30 is represented in matrix form, it can be written as $\ddot{\boldsymbol{Y}} = \ddot{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{e}$. Then, the FE estimator for $\boldsymbol{\gamma}$ is

$$\hat{\boldsymbol{\gamma}}_{FE} = (\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{X}})^{-1}\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{Y}} = \left(\frac{1}{N}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{X}}_j\right)^{-1}\left(\frac{1}{N}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\ddot{\boldsymbol{Y}}_j\right), \tag{2.31}$$

where the term N represents the total number of students (or level-1 units). The coefficient estimate $\hat{\boldsymbol{\gamma}}_{FE}$ for the covariate in the FE model is the within-cluster effect, which is identical to that for the cluster-mean centering approaches (Wooldridge, 2013, Chapter 14.3):

$$\hat{\boldsymbol{\gamma}}_{FE} = \hat{\boldsymbol{\gamma}}_{w}. \tag{2.32}$$

Further, the pooled OLS variance for the FE estimator is

$$Var(\hat{\boldsymbol{\gamma}}_{FE}) = \frac{1}{N}\boldsymbol{BMB}, \tag{2.33}$$

where $\boldsymbol{B} = (\frac{1}{N}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{X}})^{-1}$ and the central matrix $\boldsymbol{M}$ is

$$\boldsymbol{M} = \frac{1}{\bar{n}J}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\boldsymbol{\Omega}_j\ddot{\boldsymbol{X}}_j, \tag{2.34}$$

where $\bar{n} = \frac{1}{J}\sum_{j=1}^{J} n_j$ is the average number of level-1 unit per cluster, and J is the number of clusters. The term $\boldsymbol{\Omega}_j = E[\boldsymbol{e}_j\boldsymbol{e}_j'|\ddot{\boldsymbol{X}}_j]$ is the $n_j \times n_j$ covariance matrix of errors in cluster j.

In the FE model, CRVE can be used to obtain an asymptotically consistent estimator of the variance. CRVE controls the dependency that might remain in the hierarchical data even after taking account of cluster dependence by using the FE estimator (Cameron et al., 2011; Cameron & Miller, 2015). For example, suppose hierarchical data where students are nested within classes, and the classes are again nested within schools. When the FE model takes account of school clustering effects, the CRVE can control the remaining dependency among classes (within schools). As a result, CRVE corrects the underestimated SE in the FE model when a source of potential clustering is ignored (McNeish et al., 2017).

When CRVE is implemented, $\boldsymbol{\Omega}_j$ in the Equation 2.34 is estimated by using $\hat{\boldsymbol{e}}_j\hat{\boldsymbol{e}}_j'$, where $\hat{\boldsymbol{e}}_j = \ddot{\boldsymbol{Y}}_j - \ddot{\boldsymbol{X}}_j\hat{\boldsymbol{\gamma}}_{FE}$. Thus, the variance estimator with CRVE is

$$Var^{CR}(\hat{\boldsymbol{\gamma}}_{FE}) = \frac{1}{\bar{n}J}\boldsymbol{B}\hat{\boldsymbol{M}}^{\boldsymbol{CR}}\boldsymbol{B}, \tag{2.35}$$

where the estimated central matrix is

$$\hat{\boldsymbol{M}}^{CR} = \frac{1}{\bar{n}J}\sum_{j=1}^{J}\ddot{\boldsymbol{X}}_j'\hat{\boldsymbol{e}}_j\hat{\boldsymbol{e}}_j'\ddot{\boldsymbol{X}}_j. \tag{2.36}$$

As the number of clusters, J, increases to infinity, the estimated central matrix of CRVE, $\hat{\boldsymbol{M}}^{CR}$, converges to the true $\boldsymbol{M}$ and provides a consistent estimate of the covariance matrix for the OLS estimator (Cameron et al., 2011). Also, note that CRVE does not affect the coefficient estimates of the covariates but only adjusts the SE estimates of the coefficients.

The FE model using CRVE (FE-CRVE) provides researchers several benefits. First, because between-cluster effects are eliminated by FE estimators, the FE model removes the correlation between the covariates at any level and the random effects. In other words, the FE model does not require the exogeneity assumption for u_{0j} and requires fewer assumptions than when estimating the HLM. If the covariate and random effects are correlated in hierarchical data, the FE model offers a more efficient estimator than HLM. Specifically, the FE estimator provides more unbiased and consistent coefficient estimates for the level-1 covariate compared to HLM with uncentered or grand-mean centered covariates (Wooldridge, 2002, Chapter 10.5; Greene, 2018, Chapter 11.4; Gardiner et al., 2009; Kennedy, 2008, Chapter 18.3).

Next, similar to the cluster-mean centering, FE-CRVE provides the within-cluster effect, γ_w, of the level-1 covariate X_{ij} (Allison, 2009; Hamaker & Muthén, 2020; Wooldridge, 2013, Chapter 14.3). In this case, the benefit to using the FE model is that it estimates the same within-cluster effect of the covariates while offering a more simple model that focuses only on level-1 covariates. Using the FE model, researchers do not need to be concerned about the level-2 variance component and covariates. Finally, the CRVE can be easily implemented within the FE model using `R` packages, such as the `lfe` (Gaure, 2013a; Gaure, 2014) and the `fixest` packages (Bergé, 2018), which provide researchers with feasible options for handling hierarchical data.

However, the FE model has a few drawbacks as compared to HLM. First, the FE model does not allow the modeling of cluster-level covariates because the FE estimator specifically removes cluster variability. Cluster-level covariates such as W_j (e.g., school resources) cannot be included in the FE model. Second, the level-2 variance component is not estimated in the FE model. Third, if there is no correlation between covariates and random effects (i.e., when the exogeneity assumption is met), the FE model provides a less efficient coefficient estimator than HLM (Antonakis et al., 2021; Wooldridge, 2013, Chapter 14.3). Because the FE estimator is the same as the within-cluster effect in the cluster-mean centering (i.e., $\hat{\gamma}_{FE} = \hat{\gamma}_w$), it indicates the estimator from the FE model and cluster-mean centering is less efficient than the estimator from the uncentered or grand-mean-centered HLM model when the exogeneity assumption is met. For more information about the difference between HLM and FE models, see Bell and Jones (2015), McNeish and Stapleton (2016), and Wooldridge (2013, Chapter 14.2).

2.3.4 Previous Research

Several studies suggested alternatives for handling endogeneity in hierarchical data. Allison (2009), McNeish and Kelley (2019), and Raudenbush and Bryk (2002) discussed the cluster-mean centering approach, including the within-RE model and the CRE model, that provides the same within-cluster effects to the FE estimates. They emphasized that the cluster-mean centering approach retains the benefits of HLM, such as estimating the level-2 random effects variance component and incorporating the level-2 predictor in the model while obtaining the estimates from the FE estimate.

Hamaker and Muthén (2020) examined different centering methods in HLM. Based on the simulated data, they demonstrated that different slope estimates are provided depending on whether the cluster-mean centering method is used. Hamaker and Muthén (2020) also revealed that in longitudinal data, the coverage rate for the within-cluster effect of the covariate was inflated when there were only a small number of level-1 units per cluster (in this case, time points per sample), such as 4. In contrast, the coverage rate for the between-cluster effects of the covariate was lower when the number of time points per sample was 40 compared to when it was 4, while keeping the total sample size fixed at 100 persons.

Antonakis et al. (2021) conducted a Monte Carlo simulation study and showed the performance of several estimation methods, including the FE model using generalized linear squares (GLS) estimation, RE models using grand-mean centering and cluster-mean centering (within-RE model), and CRE models, which were estimated with ML and GLS, respectively. Their study concluded that when the exogeneity assumption does not hold, the RE model with grand-mean centering showed biased and inconsistent coefficient estimates of level-1 covariates. They also examined the performance of the level-2 covariates' coefficient estimates in RE and CRE models and recommended the CRE approaches to obtain a consistent estimate. However, Antonakis et al. (2021) only focused on the models that handle purely hierarchical data, not extended to complex nested data structures like cross-classified data.

Previous systematic reviews have examined how the endogeneity problem has been implemented in applied educational research. Dedrick et al. (2009) and Luo et al. (2021) summarised the practice of centering and testing assumptions in HLM across a twenty-year timeframe. Dedrick et al. (2009) provided a systematic review of HLM applications in education and related fields between 1992 and 2002. In their result, both grand-mean and cluster-mean centering were not commonly used in the literature. Only 22 studies (22%) used grand-mean centering, and 11 studies (11%) used cluster-mean centering for their level-1 covariates. Regarding testing assumptions, Dedrick et al. (2009) found that HLM assumptions were rarely mentioned, and the tenability of assumptions was seldom evaluated. The exogeneity assumptions were not examined.

Luo et al. (2021) reviewed HLM practices over the next decade, published between 2009 and 2018, using the same research journals as Dedrick et al. (2009). In the results of the comparison of trends over time, Luo et al. (2021) showed improvement in terms of the use of centering. The percentage of studies using centering increased for both grand-mean centering (34%) and cluster-mean centering (28.4%) for the level-1 covariates. However, although the overall trend improved, studies often did not discuss how centering was handled. Further, the authors found that HLM assumptions were still rarely discussed. They pointed out that testing assumptions was often an omitted step in HLM applications.

Focusing on the trends of exogeneity assumption testing in HLM, Antonakis et al. (2021) conducted a systematic review using research in the management and applied psychology fields. The authors used seven journals published between 2016 and 2017 and selected a sample of 204 articles. They found that 8 articles (4%) appropriately tested whether the exogeneity assumption held. In the rest of the studies, 98 articles (48%) used the approach that did not require this assumption (cluster-mean centering or FE estimator), and 96 articles (47%) made this assumption but did not test it in the study. However, the authors did not explicitly specify whether the selected studies used cluster-mean centering or the FE model estimator. While these two estimators are theoretically equivalent, it remains uncertain which approach the researchers preferred to use to address the endogeneity issue.

According to the three systematic reviews examined, the exogeneity assumption was not typically tested nor well handled in the applied HLM literature. Even when the exogeneity assumption does not hold, researchers do not appear to have applied centering effectively (Antonakis et al., 2021). Also, the proportion of applied studies that reported using cluster-mean centering has increased in the last decade although testing of exogeneity is insufficient. However, it is unclear why researchers chose the relevant centering method. For example, even if the researchers properly tested the exogeneity assumption, they may have used centering solely to make it easier to interpret the coefficients. In other words, the model assumption and choice of centering method as a solution seem at best to be weakly connected in applied HLM studies.

2.4 Alternatives in Cross-Classified Data

Endogeneity problems are not limited to hierarchical data. In cross-classified data, covariates can also be correlated with random effects, as in purely hierarchical data. As alternatives to solving the endogeneity problem of cross-classified data, this section describes centering methods in CCREM and two-way FE-CRVE.

I first review CCREM with not centering and grand-mean centering. Then, I discuss cluster-mean centering, which is most closely related to exogeneity assumptions. In CCREM, only a within-RE model for estimating within-cluster effects has been proposed in terms of centering options. Therefore, I propose the CRE model that provides within- and between-cluster or contextual effects for the level-1 covariate in a two-level CCREM. I also suggest cell-mean centering that takes account of interaction effects between the dimension within a covariate. Further, the FE approach can be applied to cross-classified data. I outline two-way FE-CRVE and also propose an FE-RE hybrid model, which uses fixed effects and random effects for each of the clustering dimensions, respectively. Lastly, I describe previous research that addressed potential violations of the exogeneity assumption through the use of level-1 covariate centering approaches.

CCREM with an uncentered level-1 covariate uses the raw value of the covariates without any modification. According to Raudenbush (2009), the CCREM equation can be represented in matrix form as follows:

$$\begin{gathered}\boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X\gamma} + \boldsymbol{Rb} + \boldsymbol{Cc} + \boldsymbol{e}, \\ \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}),\end{gathered} \tag{2.37}$$

where $\boldsymbol{Y}$ is a $N \times 1$ vector of observed outcomes, $\boldsymbol{1}$ is a $N \times 1$ column vector with all elements equal to 1, and γ_{000} is a fixed intercept. $\boldsymbol{X}$ is a $N \times q$ known design matrix where q is the number of covariates, and $\boldsymbol{\gamma}$ represents $q \times 1$ vector of unknown regression coefficients. $\boldsymbol{R}$ is a $N \times J$ known design matrix that includes indicators that assign $\boldsymbol{b}$ to the appropriate row, and $\boldsymbol{C}$ is a $N \times K$ known design matrix that assigns $\boldsymbol{c}$ to the appropriate columns, where $\boldsymbol{b}$ and $\boldsymbol{c}$ are $J \times 1$ and $K \times 1$ vectors of unknown random effects from each dimension, respectively. The random effects $\boldsymbol{b}$ and $\boldsymbol{c}$ are assumed to follow normal distributions with means of 0 and variances of τ_{j00} and τ_{k00}. Finally, the term $\boldsymbol{e}$ is a $N \times 1$ vector of unknown random effects, which is assumed to follow a normal distribution with a mean of 0 and variance of $\sigma^2\boldsymbol{I}$.

An interaction term, d_{0jk} (in matrix form, $\boldsymbol{d}$), as depicted in Equation 2.13, can be included in the model to capture the interaction effects between dimensions. However, to maintain consistency with Raudenbush (2009)'s notation and highlight the conceptual difference with the later introduced cell-mean centering approach, I initially focus on equations where the interaction term is omitted and assumed to be zero. In a subsequent section, the concept of cell-mean centering will be explored, which takes into account the interaction effects between clustering dimensions as well as the random interaction effect.

The intercept of CCREM with uncentered covariates is interpreted as the expected outcome value in a case where the values of each covariate are zero. However, when the zero values of the covariate are not reasonable, the interpretation of the intercept would not be meaningful. For example, if the covariate is a student's age and there are no other covariates, the intercept of the CCREM is interpreted as the expected academic achievement of a zero-year-old student. This interpretation is unrealistic and demonstrates the need for different types of centering.

The analysis of cross-classified data using CCREM involves considering multiple sources of effect for a covariate. Suppose a student is cross-classified by school and neighborhood. When a researcher analyzes the impact of a student's reading ability (i.e., level-1 covariate) on academic achievement in cross-classified data using CCREM, the resulting estimate of the covariate effect combines multiple main sources of effect other than the within-cluster effect: the between-school effect, the between-neighborhood effect, and the school by neighborhood (interaction) effect (Raudenbush & Bryk, 2002, Chapter 12).

The between-school effect indicates the expected difference in average academic achievement between two schools that differ by one unit in average reading ability. In other words, the effect is the difference in the predicted academic achievement of students from the same neighborhood but schools with different average reading abilities. Similarly, the between-neighborhood effect indicates the extent of the difference in the predicted average academic achievement between students attending the same school but living in neighborhoods with different average reading ability. For example, if this effect is positive, students living in a neighborhood with higher average reading ability may have higher average achievement scores than students at the same school but who come from a neighborhood with lower average reading ability.

The last type of effect refers to a potential interaction effect between schools and neighborhoods. When there is an interaction effect, the average effect of, say, school average reading ability on student academic achievement depends on neighborhood average reading ability. In this example then, the association between school average reading ability and achievement for students from a neighborhood with a certain average reading ability would be predicted to be a different value (given an interaction effect) should the students from that school be from a neighborhood with a different average reading ability. However, this last source of variability in the relationship between the level-1 covariate, here reading ability, and the outcome, academic achievement has received little attention in previous literature (e.g., Raudenbush, 2009).

According to the HLM literature, the most efficient estimates of the slope parameter for a covariate are obtained when the within- and between-cluster effects are assumed equal (Raudenbush & Bryk, 2002, Chapter 5). However, in empirical hierarchical data, previous literature reported that within- and between-cluster effects might not be identical (Palta & Seplaki, 2002). As such, when the within- and between-cluster effects are different in cross-classified data, it can be speculated that any of the covariate effects that are not considered in the estimating model may be captured in the relevant random effects. In that case, the corresponding covariate effects are correlated to random effects, and the CCREM exogeneity assumption would not be met, as in the case of hierarchical data (e.g., Bell & Jones, 2015; Palta & Seplaki, 2002). Thus, considering the consequences of the uncentered covariates in hierarchical data, the CCREM with uncentered covariates might also leave negative consequences that can result from the violation of the exogeneity assumption and can provide inefficient estimators.

2.4.1 Grand-Mean Centering

The use of grand-mean centering for covariates with cross-classified data context has not been explicitly discussed in the literature. However, the logic of grand-mean centering in CCREM should be similar to its use in HLM. As with the HLM, grand-mean centering in the CCREM is:

$$\begin{gathered}\boldsymbol{Y} = \mathbf{1}\gamma_{000} + \boldsymbol{X}_{grand}\boldsymbol{\gamma} + \boldsymbol{R}\boldsymbol{b} + \boldsymbol{C}\boldsymbol{c} + \boldsymbol{e}, \\ \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}),\end{gathered} \tag{2.38}$$

where $\boldsymbol{X}_{grand}$ is calculated by subtracting the overall mean of the covariate across all observations from each relevant covariate in $\boldsymbol{X}$. As mentioned, the random interaction effect d_{0jk} (or in matrix form, $\boldsymbol{d}$) can be included in the grand-mean centering method but omitted here to align with the uncentered CCREM.

Using grand-mean centering, only the intercept's estimate (and its standard error) will change because grand-mean centering is a simple linear transformation of the uncentered level-1 covariate. The intercept γ_{000} now represents the expected academic achievement score for those at the mean on each covariate in $\boldsymbol{X}_{grand}$. For example, if the covariate in $\boldsymbol{X}$ is the students' age, γ_{000} represents the overall average academic achievement for students at the grand mean age. Thus, grand-mean centering provides a meaningful intercept to interpret, compared to the model with an uncentered level-1 covariate.

However, with grand-mean centering, the slope parameter(s), the SE of the intercept, and the SE of the slope parameter estimate(s) are the same as they would be for the model with an uncentered covariate (Kreft et al., 1995). The slope coefficient estimate of the grand-mean-centered covariate still represents the pooled effect of the covariate and cannot be interpreted as within- and between-cluster effects separately. It is important to note that grand-mean centering does not solve endogeneity problems that result from a correlation between covariates and random effects. Therefore, cluster-mean centering in CCREM could be used to solve endogeneity issues, as in the HLM.

2.4.2 Cluster-Mean Centering

The cluster-mean centering approaches used with the CCREM can be categorized into the within-RE model and the CRE model, depending on the effect to be estimated. I again omitted random interaction effects in these approaches to align cluster-mean centering with Raudenbush (2009)'s notation and the uncentered and grand-mean-centered CCREM. The random interaction effects will be discussed in the cell-mean centering section.

Within-Random Effects Model

Raudenbush (2009) proposed a general model of adaptive centering that is applicable to unbalanced, cross-classified designs. The within-RE model is

$$\boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X}_{cluster}\boldsymbol{\gamma} + \boldsymbol{Rb} + \boldsymbol{Cc} + \boldsymbol{e}, \tag{2.39}$$

$$\boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}), \tag{2.40}$$

where the design matrix of cluster-mean-centered covariates is calculated as

$$\begin{aligned} \boldsymbol{X}_{cluster} &= \boldsymbol{X} - \boldsymbol{A}[\boldsymbol{A}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{A}]^{-1}\boldsymbol{A}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}, \\ &= \boldsymbol{X} - \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'\boldsymbol{X}, \end{aligned} \tag{2.41}$$

when $\boldsymbol{A} = (\boldsymbol{R} \;\; \boldsymbol{C})$. Rearranging Equation 2.41 shows that it requires

$$\boldsymbol{X}'_{cluster}\boldsymbol{A} = 0, \tag{2.42}$$

$$(\boldsymbol{X}'_{cluster}\boldsymbol{R} \;\; \boldsymbol{X}'_{cluster}\boldsymbol{C}) = 0, \tag{2.43}$$

to implement cluster-mean-centered covariates. Specifically, by assuming a homoscedastic variance, $\sigma^2\boldsymbol{I}$, $\boldsymbol{X}$ can be regressed on $\boldsymbol{C}$ and $\boldsymbol{R}$ using OLS, and the resulting residuals can then be used as the cluster-mean-centered covariate $\boldsymbol{X}_{cluster}$.

This technique of adaptive centering is based on the Frisch-Waugh-Lovell

(FWL) theorem, which allows for estimating the relevant regression coefficient in a model with multiple covariates when the focus is only on γ of the specific covariate (Lovell, 2008; Davidson & MacKinnon, 2004, Chapter 4.2). Consider $\boldsymbol{X}$, $\boldsymbol{R}$, and $\boldsymbol{C}$ are the independent variables in a multiple regression model with a dependent variable, $\boldsymbol{Y}$. When the focus is on the coefficient of $\boldsymbol{X}$, the FWL theorem calculates the regression coefficients using the following steps. First, regress $\boldsymbol{Y}$ on $\boldsymbol{R}$ and $\boldsymbol{C}$, and obtain the residuals from this regression. Then, regress $\boldsymbol{X}$ on $\boldsymbol{R}$ and $\boldsymbol{C}$, and obtain the residuals from this second regression. Finally, regress the residualized $\boldsymbol{Y}$ from the first regression on the residualized $\boldsymbol{X}$ values from the second regression. The resulting OLS estimator gives the coefficient estimate of $\boldsymbol{X}$. Alternatively, when focusing on estimating the regression coefficient, one can first regress $\boldsymbol{X}$ on $\boldsymbol{R}$ and $\boldsymbol{C}$ and then regress $\boldsymbol{Y}$ on the resulting residualized $\boldsymbol{X}$ without using the residualized $\boldsymbol{Y}$. The result is equivalent to estimating the entire model using ordinary least squares.

Similarly, the adaptive centering for the within-RE model involves taking the residuals from a regression of the covariate on a set of indicators for the cluster. This technique makes the covariate orthogonal and independent of the full set of cluster indicators, ensuring no correlation can exist between the covariate and the random effects (Raudenbush, 2009). Given inclusion of cluster-mean-centered covariates, the ML estimator for $\boldsymbol{\gamma}$ is

$$\begin{aligned}\hat{\gamma} &= [\boldsymbol{X}'_{cluster}(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}_{cluster}]^{-1}\boldsymbol{X}'_{cluster}(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{Y} \\ &= (\boldsymbol{X}'_{cluster}\boldsymbol{X}_{cluster})^{-1}\boldsymbol{X}'_{cluster}\boldsymbol{Y},\end{aligned} \tag{2.44}$$

and the corresponding variance estimator is

$$Var(\hat{\gamma}) = [\boldsymbol{X}'_{cluster}(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}_{cluster}]^{-1}, \tag{2.45}$$

where the conditional expectation of the estimator is equal to the true value of the population parameter of the coefficient, $E(\hat{\boldsymbol{\gamma}}|\boldsymbol{X}_{cluster}) = \boldsymbol{\gamma}$, even when the exogeneity assumption is violated. In other words, the within-RE model for CCREM effectively handles the correlation between the random effects and the covariates by reducing cluster variability in either dimension from the covariates. Further, this ML estimator for $\boldsymbol{\gamma}$ is exactly equivalent to the FE estimator (Raudenbush, 2009). However, the estimated coefficient from the adaptive centering may be less efficient (i.e., less precise) than the CCREM with uncentered or grand-mean-centered covariates when all the CCREM assumptions, including the exogeneity assumption, are satisfied.

Compared to the pooled effects of the covariate estimated without any cluster-mean centering or grand-mean centering, the estimator $\hat{\boldsymbol{\gamma}}$ in the within-RE model captures the within-cluster effects. Directly obtaining within-cluster effects of the covariate is useful in distinguishing the individual-level effects from the overall influence of the covariate. Other parameters, such as the intercept and variance components of the within-RE model, differ from those for a model with an uncentered or grand-mean-centered level-1 covariate. The new intercept is an intercept controlling for the cluster-mean-centered covariate. In addition, the variance components are computed by considering the removed between-cluster variances.

Correlated Random Effects model

A potential limitation of the within-RE model is that it does not provide information about between-cluster or contextual effects. Such effects are worth estimating if researchers are interested in how much each cluster dimension effect is associated with the relationship between a student's predictor and outcome. For example, Pedersen et al. (2018) investigated the influence of school and neighborhood clustering dimensions on alcohol consumption in urban adolescents, respectively. The models estimated included uncentered predictors of parents' and friends' use of alcohol. In this case, researchers might produce more nuanced results for the study if they had examined the between-cluster or contextual effects of school or neighborhood clustering on the relationship between parental or peer alcohol use and alcohol use.

In this section, I propose the CRE model for the CCREM by extending the within-RE model in CCREM. Compared to the within-RE model, the CRE model includes each dimension's cluster means as well as the cluster-mean-centered covariate and thus provides estimates of the within- and between-cluster effects:

$$\begin{gathered}\boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X}_{cluster}\boldsymbol{\gamma}_w + \bar{\boldsymbol{X}}_J\boldsymbol{\gamma}_{b,J} + \bar{\boldsymbol{X}}_K\boldsymbol{\gamma}_{b,K} + \boldsymbol{Rb} + \boldsymbol{Cc} + \boldsymbol{e}, \\ \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}),\end{gathered} \tag{2.46}$$

where $\boldsymbol{X}_{cluster}$ contains the cluster-mean-centered covariates using adaptive centering as in the within-RE model, $\bar{\boldsymbol{X}}_J$ and $\bar{\boldsymbol{X}}_K$ are design matrix of cluster means for each dimension, and $\boldsymbol{\gamma}_{b,J}$ and $\boldsymbol{\gamma}_{b,K}$ are the vectors of unknown regression coefficients indicating between-cluster effects for each dimension directly. Of the between-cluster effects, the vector of coefficients $\boldsymbol{\gamma}_{b,J}$ represents the predicted difference between two schools' average academic achievement whose average reading scores differ by one while controlling for the neighborhoods in which associated students live. Conversely, the coefficient in the vector $\boldsymbol{\gamma}_{b,K}$ is the predicted difference in the average academic attainment between two neighborhoods whose average reading scores differ by 1 unit while controlling for schools attended by students in the different neighborhoods.

The variance components of the CRE model differ from the within-RE model due to the inclusion of the additional covariates, the cluster means. Specifically, the variance components in the CRE models may be smaller than those in the within-RE model because the cluster means explain the between-cluster variability for each clustering dimension in the model.

Researchers may be more interested in the contextual effects that reflect the association between higher-level cluster effects of the covariate and level-1 unit (e.g., student) outcomes. Contextual effects in HLM can be explicitly estimated through models that include uncentered covariates and cluster means. While the coefficient for the uncentered covariates indicates the within-cluster effect, the coefficient of the cluster means indicates the contextual effect of the covariates.

Similarly, when the cluster means per each clustering dimension are included in the CCREM with the uncentered covariates, the coefficient estimates of the cluster means indicates the contextual effects of the covariates. The CCREM's contextual effect differs from that in the HLM in that it is the contextual effect for the unique clustering dimension. For example, the contextual effect of the school dimension captures the predicted difference in the academic achievement of two students from a school whose average reading scores differ by one unit while controlling for the neighborhood dimension ($\gamma_{c,J} = \gamma_{b,J} - \gamma_w$). On the other hand, the contextual effect for the neighborhood dimension indicates the difference in academic achievement for two students who go to the same school but from different neighborhoods where the average reading scores for the neighborhoods differ by one unit ($\gamma_{c,K} = \gamma_{b,K} - \gamma_w$).

Latent Mean Centering

It is worth noting that the cluster-mean centering discussed in this study uses the observed cluster means, which have a certain limitation. The observed cluster mean is calculated by averaging the covariate values of individuals sampled from the population within a cluster. Unless the entire population of the cluster is used, the observed cluster mean is an approximation of the true cluster mean. Thus, when the sampling ratio, the proportion of sampled individuals (level-1) within a cluster, is small, the observed cluster means may have a bias due to measurement error and be an unreliable estimate (O'Brien, 1990; Raudenbush et al., 1991). In such cases, the contextual effect may be underestimated, and the level-2 variance component may be inflated while the within-cluster effect remains an unbiased estimator (Grilli & Rampichini, 2011; Lüdtke et al., 2008; Shin & Raudenbush, 2010). This issue could also be relevant in unbalanced cross-classified data because fewer individuals in some clusters per dimension can lead to inconsistent estimates of the true mean.

In previous literature, researchers have proposed latent mean centering as a potential solution to address the limitations associated with using observed cluster means in cluster-mean centering (Asparouhov & Muthén, 2019; Croon & van Veldhoven, 2007; Preacher et al., 2010). Regarding the aggregating type of the covariate, latent mean centering is particularly suitable for *reflective* indicators that measure latent constructs, such as school climate, as assessed through students' evaluations (Grilli & Rampichini, 2011; Lüdtke et al., 2008). For a *formative* indicator, such as an average of students' test scores within a school, latent mean centering is still appropriate when the number of units within a cluster is infinite. On the other hand, observed mean centering is recommended for the formative indicator where the population is known and there are a finite number of units per cluster.

Lüdtke et al. (2008) primarily focused on comparing latent mean centering and observed mean centering in the random intercept model, and Asparouhov and Muthén (2019) further extended this argument to the random slope model using Bayesian estimation. Asparouhov and Muthén (2019) and Lüdtke et al. (2008) demonstrated that latent mean centering provides a relatively unbiased estimator compared to the observed mean centering, specifically when the sampling ratio is insufficient as 20%. Also, latent mean centering provides an asymptotically consistent estimator of contextual effects when there is a large number of clusters. The number of units per cluster and ICC are other factors that affect the performance of latent mean centering (Asparouhov & Muthén, 2019). For example, in practical settings with a small sampling ratio or small ICC, the latent mean centering exhibits substantial sampling variability. Thus, latent mean centering was clearly recommended for specific conditions, such as when the sampling ratio was small, and the number of clusters and the ICC were sufficiently large.

Considering the characteristics of latent mean centering, however, this study focused on observed mean centering instead of latent mean centering for several reasons. Firstly, latent mean centering outperforms observed mean centering in terms of an estimator of contextual (or between-cluster) effects. However, for the within-cluster effects, both approaches are known to provide unbiased, similar estimates.

Considering this study focuses on the performance of within-cluster and between-cluster effects, comparing the two approaches on the within-cluster effect would not be meaningful. Thus, further research with a particular focus on contextual effects is needed to investigate the comparison between latent mean centering and observed mean centering.

Moreover, though latent mean centering would provide an unbiased estimator of the contextual effect, its sampling variability largely depended on the data constellation (e.g., sampling ratio, the number of clusters, and ICC). The diverse data conditions that influence the performance of latent mean centering can complicate the interpretation of simulation study results. For example, including latent mean centering might require a separate topic concerning the sampling ratio. Lastly, latent mean centering on cross-classified data requires further research to implement in practice. While latent mean centering for hierarchical data has been extensively discussed, its applicability and implementation in cross-classified data settings have not been explored yet. Additionally, since observed mean centering is not widely used in CCREM, this study primarily focused on evaluating the performance of alternative methods in addressing the endogeneity problem utilizing observed mean centering. Future studies can explore and compare latent mean clustering and observed mean centering within the context of cross-classified data structures.

2.4.3 Cell-Mean Centering

Within-Cell Random Effects Model

Previous research suggested modeling random interaction effects for the cross-classified dimensions in CCREM estimation, which is a distinguishing characteristic of cross-classified data (Raudenbush & Bryk, 2002; Shi et al., 2010). In the same context as interaction effects between random effects, cell-interaction effects may also arise between the cluster effects of covariates. However, the cluster-mean centering approaches in the literature so far have not accounted for these interaction effects.

To address this, it would be useful to incorporate the cell-interaction effects of the covariate when employing cluster-mean centering. This approach satisfies the exogeneity assumption while considering the covariate's cell-interaction effects. In this section, I introduce the concept of cell-mean centering and its application.

Cell-mean centering uses a cell-mean-centered covariate, which is calculated by subtracting the mean of cells where each dimension of the cross-classified data intersects from the covariate,

$$X_{cell} = X_{i(jk)} - \bar{X}_{jk}, \tag{2.47}$$

where $\bar{X}_{jk}$ represents the mean of the combination cell jk of school j and neighborhood k. This cell-mean-centered covariate accounts for the cell-interaction effects of a covariate inside the model (see the detail in the Appendix 6.1).

Similar to cluster-mean centering, adaptive centering can also be used to perform cell-mean centering on unbalanced data. In this instance, a general model of the within-cell RE model using adaptive centering is:

$$\begin{gathered} \boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X}_{cell}\boldsymbol{\gamma} + \boldsymbol{Rb} + \boldsymbol{Cc} + \boldsymbol{Td} + \boldsymbol{e}, \\ \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{d} \sim N(0, \tau_{(jk)00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}) \end{gathered} \tag{2.48}$$

where the design matrix of cell-mean-centered covariate is

$$\begin{aligned} \boldsymbol{X}_{cell} &= \boldsymbol{X} - \boldsymbol{T}[\boldsymbol{T}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{T}]^{-1}\boldsymbol{T}'(\sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}, \\ &= \boldsymbol{X} - \boldsymbol{T}[\boldsymbol{T}'\boldsymbol{T}]^{-1}\boldsymbol{T}'\boldsymbol{X}, \end{aligned} \tag{2.49}$$

where $\boldsymbol{T}$ is an $N \times L$ known design matrix with the indicators that assign an $L \times 1$ vector of random interaction effects, $\boldsymbol{d}$, to the appropriate cells. Here, L is the number of unique interactions that is less than or equal to the total number of possible combinations (i.e., $L \leq J \times K$) because not all combinations of school and neighborhood might be observed in unbalanced design data. Cell-mean centering uses $\boldsymbol{T}$ that takes into account all between-cluster effects and cell-interactions effect for calculating the cell-mean-centered covariate as $(\boldsymbol{R} \;\; \boldsymbol{C})$ are a linear combination of the interactions $\boldsymbol{T}$.

Similar to conducting cluster-mean centering in the within-RE model, rearranging Equation 2.49 requires

$$\boldsymbol{X}'_{cell}\boldsymbol{T} = 0, \tag{2.50}$$

which shows how to implement cell-mean centering using the FWL theorem. Specifically, $\boldsymbol{X}$ should be first regressed on $\boldsymbol{T}$, while assuming the homoscedastic variance, $\sigma^2\boldsymbol{I}$. Then, the extracted residual from the regression can be incorporated into the model as a cell-mean-centered covariate, $\boldsymbol{X}_{cell}$. This cell-mean-centered covariate is equivalent to subtracting the cell mean from the original covariate, as demonstrated in Equation 2.47.

The within-cell RE model is expected to be robust to violation of the exogeneity assumption, similar to what is accomplished using cluster-mean centering by eliminating between-cluster variability. Compared to cluster-mean centering, the within-cluster effect in cell-mean centering can differ from that in cluster-mean centering because it additionally controls the cell-interaction variability. In this sense, the within-cluster effect in this model can be interpreted as the predicted difference in student academic achievement for two students with a one-unit difference in reading ability who share the same neighborhood and school. However, it has yet to be studied how the coefficients estimated when using cell-mean centering differ from those when using cluster-mean centering.

Correlated-Cell Random Effects model

Similar to the CRE model with cluster-mean centering, the cluster mean per clustering dimension and the cell mean can be included in the correlated-cell RE model as control variables. The correlated-cell RE model is

$$\begin{aligned} \boldsymbol{Y} =& \boldsymbol{1}\gamma_{000} + \boldsymbol{X}_{cell}\boldsymbol{\gamma}_{w} + \bar{\boldsymbol{X}}_{J}\boldsymbol{\gamma}_{b,J} + \bar{\boldsymbol{X}}_{K}\boldsymbol{\gamma}_{b,K} + \bar{\boldsymbol{X}}_{JK}\boldsymbol{\gamma}_{b,JK} + \boldsymbol{Rb} + \boldsymbol{Cc} + \boldsymbol{Td} + \boldsymbol{e}, \\ & \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{d} \sim N(0, \tau_{(jk)00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^{2}\boldsymbol{I}), \end{aligned} \tag{2.51}$$

where $\boldsymbol{X}_{cell}$ is the adaptive cell-mean-centered covariate using Equation 2.49 as in the within-cell RE model, $\bar{\boldsymbol{X}}_{J}$, and $\bar{\boldsymbol{X}}_{K}$ are the design matrices of cluster-means per dimension and $\boldsymbol{\gamma}_{b,J}$, and $\boldsymbol{\gamma}_{b,K}$ are the corresponding vectors of the between-cluster effects. Likewise, $\bar{\boldsymbol{X}}_{JK}$ is the design matrix of adaptively centered cell means for the vector of the cell-interaction effect, $\boldsymbol{\gamma}_{b,JK}$. Using the FWL theorem, the adaptively centered cell mean is obtained by extracting the residuals of regressing the cell mean on $(\boldsymbol{R} \;\; \boldsymbol{C})$. Using this method, multicollinearity that may arise between the two cluster means and cell means can be solved.

The correlated-cell RE model estimates identical within-cluster effects to those estimated using the within-cell RE model. Further, this model includes cell means as well as cluster means, allowing for the estimation of the between-cluster and cell-interaction effects. The between-cluster and cell-interaction effects are orthogonal to the estimated within-cluster effect because the adaptive centering approach was used to partial out between-cluster effects and the cell-interaction effect from the covariate variance. In terms of interpretation, the between-cluster effect represents the predicted difference in average academic achievement between two clusters with a one-unit difference in average reading scores while controlling for the cell-interaction effect. Similarly, the cell-interaction effect indicates the predicted difference in average academic achievement between two cells with a one-unit difference in average reading scores while controlling for school and neighborhood effects. However, the included cell means are likely to be used as an additional control variable rather than as the main focus of the study.

2.4.4 Two-Way FE-CRVE

In the way that the one-way FE-CRVE has been an option for handling endogeneity in purely hierarchical data, the two-way FE model using CRVE offers an alternative for handling endogeneity in cross-classified data. Considering the same cross-classified data structure as in CCREM, where individuals $i \in 1, 2, ..., n_j$ are within schools $j \in 1, 2, ..., J$ and neighborhoods $k \in 1, 2, ..., K$, the two-way FE model with one covariate is

$$Y_{i(jk)} = \gamma_{100} X_{i(jk)} + \sum_{p=1}^{J} \gamma_{0p0} D_p + \sum_{q=1}^{K} \gamma_{00q} D_q + e_{i(jk)}, \tag{2.52}$$

where γ_{100} is the coefficient estimate for the covariate $X_{i(jk)}$, γ_{0p0} is a school-level intercept for the school indicator D_p, and γ_{00q} is a neighborhood-level intercept for the neighborhood indicator D_q as a dummy variable. The term $e_{i(jk)}$ is a random error for a student i nested within school j and neighborhood k. The random errors for student $i(jk)$ are assumed to be independent of random errors for students belonging to different schools and neighborhoods (i.e., $m(gh)$).

Alternatively, the matrix form of the two-way FE model can be expressed as:

$$\boldsymbol{Y} = \boldsymbol{X\gamma} + \boldsymbol{R\gamma}_j + \boldsymbol{C\gamma}_k + \boldsymbol{e}, \tag{2.53}$$

where $\boldsymbol{\gamma}_j$ is a $J \times 1$ vector of school-level intercepts associated with an $N \times J$ known design matrix, $\boldsymbol{R}$, that contains the indicators for each of the j clusters in the clustering dimension J, and $\boldsymbol{\gamma}_k$ is a $K \times 1$ vector of neighborhood-level intercepts associated with $\boldsymbol{C}$, an $N \times K$ known design matrix with the neighborhood indicators.

However, when the number of clusters is large for either dimension, it is possible to apply a generalization of the FE estimator known as the *Method of Alternating Projections* and avoid estimating a fixed effect for each cluster (Gaure, 2013b). This method projects two-dimensional fixed effects on outcome variables and covariates and derives the model's coefficient estimates without requiring complicated fixed-effects calculations. That is, it is possible to estimate coefficients efficiently even if each dimension contains numerous clusters. Consequently, the two-way FE estimator removes cluster-level variability and accounts for the correlation between cluster-level random effects and covariates for both dimensions.

Using CRVE, the FE model can manage dependencies among clusters that have not been explicitly specified. Cameron et al. (2011) and Cameron and Miller (2015) extended the one-way CRVE and suggested the two-way CRVE that can account for any dependence that remains in the errors within any cluster dimensions. The authors demonstrated that the FE model using two-way CRVE performs well for balanced cross-classified data and yields an acceptable Type I error rate for a covariate's coefficient estimate. Thompson (2011) also indicated that the FE model (panel regression) using two-way CRVE provided more accurate inferences than from use of the one-way CRVE with cross-classified data.

Given that the variance estimator of the two-way FE model is

$$Var(\hat{\boldsymbol{\gamma}}_{FE}) = \frac{1}{N}\boldsymbol{BMB}, \tag{2.54}$$

where $\boldsymbol{B} = \frac{1}{N}\sum_{j=1}^{J}(\ddot{\boldsymbol{X}}'\ddot{\boldsymbol{X}})^{-1}$ and $\ddot{\boldsymbol{X}}$ is the adaptively cluster-mean-centered covariate using FWL theorem, the central matrix $\boldsymbol{M}$ is

$$\boldsymbol{M} = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{N}\ddot{\boldsymbol{X}}_{i(jk)}\ddot{\boldsymbol{X}}'_{m(gh)} \times Cov(\boldsymbol{e}_{i(jk)}, \boldsymbol{e}_{m(gh)}), \tag{2.55}$$

where N is the total number of students.

In two-way CRVE, the covariance between errors is estimated using the products of residuals for pairs of observations (students) who share one or both dimensions (same school or same neighborhood). Thus, the estimated central matrix for the variance estimator with two-way CRVE is

$$\hat{\boldsymbol{M}}^{CR} = \frac{1}{N}\ddot{\boldsymbol{X}}'(\hat{\boldsymbol{e}}\hat{\boldsymbol{e}}' \circ \boldsymbol{S}^{JK})\ddot{\boldsymbol{X}}, \tag{2.56}$$

where the notation $\circ$ denotes element-wise multiplication, and $\boldsymbol{S}^{JK}$ is an $N \times N$ matrix in which only the entry that shares a cluster in one or both dimensions are equal to 1 and 0 otherwise. $\hat{\boldsymbol{M}}^{CR}$ is an asymptotically consistent estimator of $\boldsymbol{M}$, meaning that as the number of clusters in both dimensions approaches infinity, $\hat{\boldsymbol{M}}^{CR}$ converges to $\boldsymbol{M}$. Then, the two-way CRVE can be written as

$$\hat{\boldsymbol{V}}^{CR} = \frac{1}{N}\boldsymbol{B}\hat{\boldsymbol{M}}^{CR}\boldsymbol{B}. \tag{2.57}$$

The substantial benefit of two-way FE-CRVE is that it handles endogeneity by eliminating the between-cluster variability in cross-classified data. Therefore, when there is a correlation between the covariate and level-2 random effects for either dimension, the two-way FE-CRVE provides a less biased and more consistent estimator than the CCREM without centering or with grand-mean centering of the level-1 covariate. In addition, the two-way FE-CRVE estimates the within-cluster effects of the level-1 covariate without requiring the specification of a complex random effects model. Thus, the two-way FE-CRVE offers a useful alternative if the researcher is focused on the within-cluster effect of the level-1 covariate coefficient and if there is a possibility of endogeneity.

Note that the coefficient for the level-1 covariate in two-way FE-CRVE is the same as that in the within-RE and the CRE models in CCREM, although the SE might not be identical (Greene, 2018, Chapter 11.5.7). These results would be the same even if the exogeneity assumption is violated because all three models eliminate between-cluster variability in the level-1 covariate that might cause the endogeneity problem. In this sense, the two-way FE-CRVE has the limitation of an inability to estimate the effect of cluster-level covariates or of variance components compared to the other two models with cluster-mean centering. In contrast, the CRE models permit the estimation of both elements.

2.4.5 FE-RE Hybrid Model

I suggest another alternative model, the FE-RE hybrid model, that uses the fixed effect and random effect simultaneously. This approach is similar to the CCREM except for the way in which one of the dimensions is treated. Specifically, one of the clustering dimensions is translated into fixed effects and the other dimension is captured as random effects in the FE-RE hybrid model. For example, the model that treats the school dimension as FE and the neighborhood dimension as RE is

$$\begin{gathered} \boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{R}\boldsymbol{\gamma}_j + \boldsymbol{C}\boldsymbol{c} + \boldsymbol{e}, \\ \boldsymbol{c} \sim N(0, \tau_{k00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}), \end{gathered} \tag{2.58}$$

where $\boldsymbol{X}$ is the design matrix of uncentered covariates, $\boldsymbol{R}$ is a $N \times J$ design matrix of dummy-coded school indicator variables and $\boldsymbol{\gamma}_j$ is a $J \times 1$ vector of the corresponding school-specific intercepts. $\boldsymbol{C}$ is an $N \times K$ matrix that contains indicators to assign the random effects $\boldsymbol{c}$ to the correct neighborhood for each individual student. Because between-cluster variability is captured for the two dimensions, $\boldsymbol{\gamma}$ is the within-cluster effect. The level-1 error $\boldsymbol{e}$ is an $N \times 1$ vector with assumed homoscedastic variance, $\sigma^2\boldsymbol{I}$.

The decision to treat certain dimensions as fixed or random effects may vary depending on the research question being addressed. For example, contrary to Equation 2.58, the neighborhood dimension can be modeled as FE, and the school dimension can be treated as RE:

$$\begin{gathered} \boldsymbol{Y} = \boldsymbol{1}\gamma_{000} + \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{C}\boldsymbol{\gamma}_k + \boldsymbol{R}\boldsymbol{b} + \boldsymbol{e}, \\ \boldsymbol{b} \sim N(0, \tau_{j00}\boldsymbol{I}), \quad \boldsymbol{e} \sim N(0, \sigma^2\boldsymbol{I}), \end{gathered} \tag{2.59}$$

where $\boldsymbol{C}$ is a $N \times K$ design matrix of dummy-coded neighborhood indicators for the neighborhood dimension and γ_k is $K \times 1$ vector of the neighborhood-specific intercepts. $\boldsymbol{R}$ is a $N \times J$ design matrix of indicators to assign a $J \times 1$ vector of school random effects $\boldsymbol{b}$.

The FE-RE hybrid model has the benefit of the FE and the RE models simultaneously. One advantage is that it allows for estimating both cluster-level covariate effects and random effects for the selected clustering dimension. Also, the FE-RE hybrid model requires fewer assumptions than CCREM with an uncentered or grand-mean-centered covariate. For the clustering dimension treated as FE, independence between the covariates and the random effects is not required. In other words, even when the covariates are correlated with the random effects of the FE dimension, the use of the FE coding provides robust coefficient estimates for that dimension's level-1 covariate.

However, considering that the exogeneity assumption is still required for the dimension treated with RE, modeling FE for one dimension does not guarantee the complete elimination of the endogeneity problem. In other words, due to the clustering dimension treated with RE, the validity of inferences associated with the use of the FE-RE hybrid model is founded on more assumptions than when using the CCREM with adaptively centered covariates. Thus, with the FE-RE hybrid model, when the covariates are correlated with the random effects for the dimension treated as RE, the estimated within-cluster effect might be biased. In this instance, performing the FE-RE hybrid model in both ways might shed light on which dimensions have endogeneity. Further, as long as the two dimensions are correlated, it is necessary to explore whether this approach fully resolves all endogeneity problems.

2.4.6 Previous Research

Compared to the number of studies involved with hierarchical data, substantially fewer studies have been conducted on methods for handling endogeneity for cross-classified data. Except for Raudenbush (2009) who suggested two-way cluster-mean centering, extending the CCREM to obtain the between-cluster or contextual effects has not been discussed. Considering that the CRE model in HLM exhibited benefits for handling cluster-level covariates and including random effects compared to the FE model, it seems critical to examine the performance of the within-cluster, between-cluster, or contextual effect estimates when using various cluster-mean centering methods in CCREM. Moreover, given the possibility of endogeneity in the CCREM, it may be necessary to provide alternatives for cluster-mean-centered models to handle potential endogeneity.

While two decades of systematic reviews have assessed the applied HLM literature, only one systematic review has been conducted on cross-classified data research. Barker et al. (2020) investigated a total of 118 empirical CCREM studies with health outcomes published between 1994 and 2018, focusing on the rationale for using CCREM. However, their review did not focus on whether the CCREM assumptions were appropriately evaluated or whether the cluster-mean centering approach was used to adjust for the endogeneity. Also, their study only included studies with health outcomes.

Therefore, as conducted by Dedrick et al. (2009) and Luo et al. (2021), a systematic review is needed to examine how the assumption of exogeneity is typically addressed in CCREM research. It is also necessary to determine whether centering is used in CCREM and what type of centering is typically used. Given the lack of demonstration for how to use cluster-mean centering with the CCREM, the frequency of use of cluster-mean centering with CCREMs is expected to be lower than with the HLM.

2.5 Empirical Examples

As an illustration of the centering methods in CCREM, I used two empirical datasets from Raudenbush and Bryk (2002) and Paterson (1991). The purpose of these examples is not to investigate the impact of the variables on the outcome (education attainment) but to illustrate the results of CCREM using various centering methods, two-way FE-CRVE, and the FE-RE hybrid model. I analyzed both empirical datasets using R 4.2.1 (R Core Team, 2022). The centering methods and the FE-RE hybrid models used the lmer() function of the lme4 package (Bates et al., 2015). The two-way FE-CRVE model was analyzed using the felm() function of the lfe package (Gaure, 2013a).

2.5.1 Example 1

The data from Raudenbush and Bryk (2002) include 2,310 students from 17 schools living in 524 neighborhoods in Scotland. Table 2.1 presents the descriptive statistics for the data structure and variable information. On average, a school has around 136 students from an average of 46 different neighborhoods. The sparsity of the unbalanced dataset was 0.088, indicating the ratio of cells filled with students among all the neighborhood and school combinations.[1] I used education attainment as an outcome variable, a composite score based on two national examinations that students took during their secondary school. This outcome variable was log-transformed to follow a normal distribution and multiplied by 10 to increase the variance of the scale. I used the primary seventh-grade reading variable for the level-1 predictor. The reading variable was a standardized score measured before students started secondary school. In Raudenbush and Bryk (2002), CCREM was applied without centering on exploring the impact of a number of covariates, including verbal reasoning scores and demographic variables. In this empirical example, however, I used only reading scores as the level-1 covariate to focus on a comparison of each coefficient based on different alternatives.

Table 2.2 displays the coefficient for the covariates and their corresponding SEs from various level-1 covariate centering methods, including CCREM without center-

[1] In an unbalanced cross-classified data structure, students typically only exist in some combinations between two clustering dimensions. For example, suppose there are $J \times K$ combinations between schools 1 through J and neighborhoods 1 through K. The number of combinations filled with students might be less than $J \times K$ (i.e., $L \leq J \times K$). Thus, the number of combinations where students are cross-classified within schools and neighborhoods would be a fraction of the total number of combinations: $L/(J \times K)$, called sparsity.

Table 2.1

Descriptive Statistics of Garner and Raudenbush (1991) Data

Variable	Mean	SD	Min	Q1	Q2	Q3	Max
Students per school	136	550	22	102	136	155	286
Students per neighborhood	4	3	1	2	4	6	16
Neighborhoods per school	46	17	11	40	43	52	92
Education Attainment	0.93	10.02	-13.28	-5.81	1.58	7.35	24.15
Primary 7th-Grade Reading	-0.04	13.89	-31.87	-9.87	-0.87	9.13	28.13

ing, grand-mean centering, the within-RE model, CRE model using cluster-mean centering and cell-mean centering, two-way FE-CRVE, and FE-RE hybrid model.

Table 2.2

Comparison of Centering using Empirical Example Data 1: Raudenbush & Bryk (2002)

	Not Centering	Grand-Mean Centering	Cluster-Mean Centering		Cell-Mean Centering		FE CRVE	FE-RE Hybrid	
Model	(1)	(2)	(3) Within-RE	(4) CRE	(5) Within-RE	(6) CRE	(7)	(8) S-FE/N-RE	(9) S-RE/N-FE
Fixed Effects									
Intercept	0.978 (0.310)	0.957 (0.310)	0.699 (0.736)	0.998 (0.278)	0.701 (0.742)	0.986 (0.276)		1.574 (0.683)	1.992 (3.231)
γ_{pooled}	0.475 (0.011)	0.475 (0.011)							
γ_w			0.440 (0.013)	0.440 (0.013)	0.434 (0.014)	0.434 (0.014)	0.440 (0.013)	0.471 (0.011)	0.441 (0.013)
$\gamma_{b,J}$				0.136 (0.066)		0.127 (0.066)			
$\gamma_{b,K}$				0.553 (0.022)		0.553 (0.022)			
$\gamma_{b,JK}$						0.478 (0.036)			
Random Effects (SD)									
School Effects	1.036	1.036	2.807	0.886	2.828	0.871			0.984
Neighborhood Effects	1.796	1.796	4.675	1.589	2.864	0.004		1.815	
Interaction Effects					4.358	1.882			
Residual	7.044	7.044	7.055	7.083	7.047	7.014		7.038	6.987

Note. The value in the parentheses indicates the standard errors of the coefficients. The predictor included is the primary 7th-grade reading score. $\gamma_{b,J}$ indicates the between-cluster effect for the school dimension, $\gamma_{b,K}$ indicates the between-cluster effect for the neighborhood dimension, and $\gamma_{b,JK}$ indicates the cell-interaction effect. When the random-intercept model was conducted with the primary 7th-grade reading score as the outcome and no covariate, the IUCC for the school dimension was 9.15%, for the neighborhood dimension was 6.12%, and the cell IUCC was 3.5%.

The first two columns show the results of the uncentered covariate (Model 1) and the grand-mean-centered covariate (Model 2). The coefficient of the reading score when uncentered and grand-mean-centered was identical, as expected, ($\gamma_{pooled} =$ 0.475, SE = 0.011), which may be interpreted as the expected difference in attainment scores between two students with reading scores that differ by one. This coefficient represents the pooled effects of the within- and between-cluster effects, and these two effects cannot be distinguished.

The intercept was different between the model with and without grand-mean centering. In the model with not centering, the intercept of 0.978 (SE = 0.310) reflects the expected attainment score when all other covariates (in this case, reading score) are zero. In contrast, grand-mean centering yields a different intercept, 0.957 (SE = 0.310), which indicates the expected attainment score for those with the grand-mean reading score (-0.04, Table 2.1). However, neither model is robust when the exogeneity assumption is violated.

Cluster-mean centering resulted in different results compared to estimates from the uncentered and grand-mean-centered covariate models. First, in the within-RE model using cluster-mean centering (Model 3), the reading variable's coefficient was 0.440 (SE = 0.013), indicating the predicted difference in the students' attainment score for two students whose reading scores differ by 1 unit within the same cluster of either school or neighborhood. The coefficient of the within-RE model was smaller than the model estimates with the uncentered and grand-mean-centered covariate because it is a within-cluster effect. The intercept, coefficients for the covariates, and the random effects variance components were not as close in value to those from the uncentered and grand-mean-centered covariate models.

Next, the CRE model using cluster-mean centering (Model 4) provided the same within-cluster effect of 0.440 (SE = 0.013) for the reading variable as the within-cluster effect estimated by the within-RE model. Further, the CRE model estimates the between-cluster effects directly by incorporating the cluster means in the model.

The between-school and between-neighborhood effects were 0.136 (SE = 0.066) and 0.553 (SE = 0.022), respectively. For instance, the predicted difference between two schools' average attainment scores, whose students' average reading scores differ by 1, was 0.136 while controlling for neighborhoods. Also, the predicted difference in average attainment score between neighborhoods where students' average reading scores differ by 1 was 0.553, while controlling for schools.

In the CRE model, the contextual effects can be calculated by subtracting the within-cluster effect from the between-cluster effect for each dimension, respectively (Note this comparison just involves an explanation of the values and is not a test of statistical significance of the difference in the values). Specifically, the value of the between-school effect was notably distinct from the within-cluster effect, particularly when compared to the between-neighborhood effect. As a result, the contextual effect on the school exhibited a greater magnitude compared to the contextual effect on the neighborhood dimension. This implies that when controlling for neighborhoods, the predicted difference in attainment scores between two students from different schools, whose average reading scores differ by one unit, becomes substantial (0.136 - 0.440 = -0.304).

I also used the cell-mean centering methods that account for the model's cell-interaction effects. In the within-cell RE model (Model 5), the within-cluster effect was 0.434 (SE = 0.014). This effect represents the predicted difference in the attainment scores between two students from the same combination of school and neighborhood whose reading scores differ by 1 unit. Because the covariates were adaptively centered by cell means, the covariance matrices of the data structures differed from that in the cluster-mean centering, resulting in different variance components.

The correlated-cell RE model (Model 6) calculates the within-cluster effect as well as the between-cluster effect and the cell-interaction effect. After including the cluster means and adaptively centered cell mean as additional covariates, the within-cluster effect was 0.434 (SE = 0.014), which was identical to the within-cluster effect in that in the within-cell RE model.

The coefficients for the cluster means were 0.127 (SE = 0.066) for the between-school effect and 0.553 (SE = 0.022) for the between-neighborhood effect. Since the cell means are adaptively centered and orthogonal to the cluster means for schools and neighborhoods, the between-cluster effects of these two clustering dimensions should be the same as in the CRE model. However, the correlated-cell RE model appears to have introduced some coefficient differences because random interaction effects are included in the model compared to the CRE model.[2] Thus, the between-school effect differed slightly from the results in cluster-mean centering approaches. The cell-interaction effect was 0.478 (SE = 0.036), which reflects the predicted difference in the average attainment scores of two school-neighborhood combinations (cell), where the average reading scores differ by 1 unit. The cell-interaction effect was greater than the between-school effects but smaller than the between-neighborhood effects.

The coefficient estimated using the two-way FE-CRVE (Model 7) was 0.440, showing the within-cluster effects using the FE estimator. This coefficient estimate of 0.440 was the same as the within-cluster effects in the cluster-mean centering. However, the two-way FE-CRVE did not estimate the variance components of the random effects as in CCREM.

Finally, I conducted two FE-RE hybrid models. When modeling the school as FE and the neighborhood as RE (Model 8), the within-cluster effect estimate was 0.470 (SE = 0.011), which was larger than the results obtained by the other approaches. Given that the hybrid model only handles endogeneity in the dimension specified as FE, this difference might indicate that there is a correlation between the covariate and the neighborhood random effects. When modeling the school dimension as RE and the neighborhood as FE (Model 9), the within-cluster effect estimate was 0.441, which was still slightly greater than the within-cluster effects from the other approaches. However, the pattern of results may be idiosyncratic to this particular

[2] When the correlated cell RE model was performed without random interaction effects, the between-cluster effects of the covariates were exactly the same as in the CRE model.

dataset. Thus, I illustrated the results of the same methods using another example, the empirical data from Paterson (1991).

2.5.2 Example 2

The data analyzed by Paterson (1991) consist of 3,435 children attending 148 primary and 19 secondary schools in Scotland. I present the descriptive statistics for the data in Table 2.3. In this dataset, the average number of students per primary school was approximately 23, and the average number of students per secondary school was around 181. The number of students who enter a secondary school from the same primary school was 16 on average. The sparsity of the data design was calculated as 0.108. Here, education attainment was the outcome variable, and the verbal reasoning score was used as the level-1 predictor. The verbal reasoning score was obtained from a test given to students when they entered secondary school. As in the first example, the outcome variable was multiplied by 10 to increase the scale of variance.

Table 2.3

Descriptive Statistics of Patterson (1991) Data

Variable	Mean	SD	Min	Q1	Q2	Q3	Max
Students per Primary School	23	17	1	8	21	36	72
Students per Secondary School	181	62	92	124	175	2,346	290
Primary per Secondary School	16	6	7	13	14	18	32
Attainment	56.79	30.69	10	30	50	90	100
Verbal Reasoning	-2.20	13.29	-30	-11	-2	7	40

Table 2.4 presents the results of alternative models estimated using the empirical dataset. The first two models, not centering (Model 1) and grand-mean centering (Model 2), both resulted in the same coefficient of 1.6 for the level-1 predictor (SE = 0.028), indicating the pooled effect of the within- and between-cluster effect of the covariate. However, these models do not take into account the potential correlation between the covariate and the random effects, which can lead to invalid statistical

inferences if the exogeneity assumption is violated in the data.

To address this issue, I used the cluster-mean centering approaches (Models 3 and 4). With these models, the within-cluster effect estimates for the level-1 predictor were identical with a value of 1.560 (SE = 0.029). The CRE model (Model 4) additionally provides estimates of the between-cluster effects for the primary school dimension (1.932, SE = 0.122) and the secondary school dimension (1.111, SE = 0.291).

The cell-mean centering approaches (Models 5 and 6) extend the cluster-mean centering models by incorporating the cell-interaction effect and the random interaction effect. The within-cluster estimate under the within-cell RE model (Model 5) was 1.560 (SE = 0.030), which was identical to the estimate under the within-cell RE model, which was also 1.560 (SE = 0.029). The correlated-cell RE model (Model 6) further provided estimates of the between-cluster effects for both dimensions and the cell-interaction effect: the between-primary-school effect was 1.928 (SE = 0.121), and the between-secondary-school effect was 1.058 (SE = 0.285). These two effects were similar values to those in the CRE model (Model 4). As in the previous dataset, one of the between-cluster effects (here, the between-neighborhood effect) showed a slight difference compared to the between-cluster effects obtained from the CRE model using the cluster-mean centering. This difference is likely due to the adaptive centering of cell-mean-covariate and additional consideration of the cell-interaction and random interaction effects in the models. Finally, the cell-interaction effect was 1.551 (SE = 0.130).

Table 2.4

Comparison of Centering using Empirical Example Data: Patterson (1991)

	Not	Grand-Mean	Cluster-Mean		Cell-Mean		FE	FE-RE	
	Centering	Centering	Centering		Centering		CRVE	Hybrid	
Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
			Within-RE	CRE	Within-RE	CRE		P-FE/S-RE	P-RE/S-FE
Fixed Effects									
Intercept	59.78	56.265	54.490	62.835	54.146	62.704		61.171	61.214
	(0.670)	(0.667)	(1.967)	(1.065)	(1.840)	(1.028)		(3.776)	(2.296)
γ_{pooled}	1.600	1.600							
	(0.028)	(0.028)							
γ_w			1.560	1.560	1.560	1.560	1.560	1.565	1.588
			(0.029)	(0.029)	(0.030)	(0.030)	(0.029)	(0.028)	(0.028)
$\gamma_{b,J}$				1.932		1.928			
				(0.122)		(0.121)			
$\gamma_{b,K}$				1.111		1.058			
				(0.291)		(0.285)			
$\gamma_{b,JK}$						1.551			
						(0.130)			
Random Effects (SD)									
Primary School Effects	5.241	5.241	12.306	5.126	8.895	4.644			5.389
Secondary School Effects	1.199	1.199	6.894	2.411	6.218	2.231		3.533	
Interaction Effects					9.755	2.307			
Residual	20.627	20.627	20.619	20.722	20.811	20.713		20.560	20.613

Note. The value in the parentheses indicates the standard errors of the coefficients. The predictor included is the verbal reasoning score. $\gamma_{b,J}$ indicates the between-cluster effect for the primary school dimension, $\gamma_{b,K}$ indicates the between-cluster effect for the secondary school dimension, and $\gamma_{b,JK}$ indicates the cell-interaction effect. When the random-intercept model was conducted with the verbal reasoning score as the outcome and no covariate, the IUCC for the primary school dimension was 4.89%, for the secondary school dimension was 3.17%, and the cell IUCC was 3.88%.

The two-way FE-CRVE (Model 7) estimates the within-cluster effect using the FE estimator for both dimensions. This model estimated the within-cluster effect as 1.560 (SE = 0.029) while controlling for potential dependencies using CRVE. As expected, the within-cluster effect from the two-way FE-CRVE is the same as the results when using cluster-mean centering.

Under the first hybrid model (Model 8), the primary school dimension was specified as FE and the secondary school dimension as RE. This model estimated the within-cluster effect as 1.565 (SE = 0.028), which was slightly larger than the estimates from the other models. This suggests that the FE-RE hybrid model can provide similar estimates for the level-1 predictor's within-cluster effect under certain conditions, such as the exogeneity assumption being met in the dimensions where it was modeled as RE.

However, when the primary school dimension was modeled as RE and the secondary school dimension as FE (Model 9), the within-cluster effect was estimated at 1.588 (SE = 0.028). This discrepancy may be due to the fact that the secondary school dimension was not controlled as FE and was instead modeled as RE, which requires meeting the exogeneity assumption. This assumption may not hold if the covariates are correlated with the random effects of the secondary school dimension.

Overall, the results of the alternative approaches vary based on the data characteristics. It is uncertain to what extent the coefficients of the covariates vary across different data conditions. Thus, simulation studies are necessary to examine the performance of these alternatives under various data conditions.

2.6 Purpose of Study

Use of both HLM and CCREM involves stringent assumptions about exogeneity, meaning that there is no correlation between covariates and random effects. To avoid this assumption, cluster-mean centering has been developed for purely hierarchical data (Allison, 2009; Hamaker & Muthén, 2020). Previous studies on cluster-mean centering have thoroughly examined the characteristics and differences between within-RE and CRE models.

A two-decade systematic review conducted over two studies has shown that cluster-mean centering became more widespread in the 2010s compared to the 2000s (Dedrick et al., 2009; Luo et al., 2021). However, because there were substantially fewer cases of HLM testing exogeneity assumptions, it remains unclear whether cluster-mean centering has been employed to resolve the endogeneity problem (Antonakis et al., 2021). In contrast to the HLM, it is unknown how CCREM research has dealt with endogeneity thus far. Only one study has investigated trends in the formulation of the CCREM, focusing on examining the rationale behind use of the CCREM for handling cross-classified data (Barker et al., 2020).

Also, approaches for minimizing the impact of the exogeneity assumptions have not been well developed for the CCREM, particularly for cluster-mean centering. Only the two-way adaptive centering developed by Raudenbush (2009) provides an option to estimate a within-RE model in CCREM. Furthermore, the interaction of two dimensions cannot be ignored due to the nature of CCREM dealing with two cross-classified cluster dimensions (Shi et al., 2010). The interaction between two dimensions of the covariates might appear in cross-classified data, which might need to be considered in cluster-mean centering. In this instance, cluster-mean centering employing the cell means, a combination of two clusters from different dimensions, could be used to allow the interaction effects between two dimensions.

Lastly, FE approaches need to be examined further. Two-way FE using CRVE suggests alternatives to handle the correlation between the covariate and the random effects using FE while adjusting the remaining dependence in the errors within potential cluster dimensions (Cameron et al., 2011; Cameron & Miller, 2015). In addition, the FE-RE hybrid model that treats each dimension of cross-classified data as FE and RE is a potential alternative that has never been explored.

Thus, there remains a need to investigate how the performance of centering methods, two-way FE-CRVE, and the FE-RE hybrid model differ from each other. It should be determined whether the estimated values of the pooled effect, within-cluster effect, and between-cluster effect accurately represent their true values when the covariates are correlated with the random effects. It is anticipated that the covariate coefficient performance at each level will vary according to the characteristics of the data. To my knowledge, however, there has been no research on these alternatives. Therefore, the purpose of this study is as follows:

1. Through a systematic review, I examine how empirical studies using CCREM have addressed exogeneity assumptions. I summarize the extent to which previous studies have employed centering approaches and what type of centering has typically been used. I also use the systematic review to help gather typical characteristics of applied CCREM studies to inform the design of the simulation study I conduct.

2. Using a Monte Carlo simulation study, I examine the performance of CCREM models, including the grand-mean centering, cluster-mean centering, and cell-mean centering, and compare these to the performance of the two-way FE-CRVE and the FE-RE hybrid model. I particularly investigate the performance of pooled effect, within- and between-cluster effect estimates for each method (for the pooled and between-cluster effect, only in the applicable models). Performance criteria include relative parameter bias, absolute parameter bias, root mean square error (RMSE), and relative bias of SE. I also examine how the performance of these methods varies under diverse data generation conditions. The benefits and drawbacks of each model are assessed, compared, and discussed.

Chapter 3

Systematic Review

3.1 Methods

In the systematic review, I explored how the empirical literature using the CCREM has evaluated model estimation assumptions and employed alternatives to handle potential assumption violations. This systematic review followed the HLM-focused criteria used in Dedrick et al. (2009) and Luo et al. (2021)'s systematic reviews. In contrast to the two studies that have covered numerous aspects of HLM, I primarily focused on the recognition of the data characteristics, the CCREM assumptions, and the use of the centering approaches. In the data characteristics, I reviewed the cross-classified data structure and the CCREM specification in the literature to serve as the rationale for the conditions in the simulation study I conducted. All systematic review processes and presentation of results were based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses 2020 (PRISMA; Page et al., 2021).

3.1.1 Inclusion and Exclusion Criteria

I included research that employed CCREM in their data analyses.[1] Studies published during the last decade, 2013 and 2022, were included in the synthesis to capture the most recent trends in applied researchers' use of the CCREM (last searched date: December 31, 2022). The studies included were limited to English-language publications. If CCREM were used as part of a methodological investigation, those were included only when the methodological research illustrated an example of

[1] If the study used the multiple-membership random effects models (MMREM), which differs from the CCREM, it was excluded. Unlike CCREM, MMREM handles complex hierarchical data structures in which lower (e.g., level-1) units (e.g., students) belong to multiple higher-level units (e.g., multiple schools).

CCREM using empirical data. Further, I excluded the applied studies that used the CCREM for SEM, meta-analysis, or item response theory (IRT) models.

3.1.2 Search Strategy and Screening

Considering the limited number of CCREM studies compared to HLM, it may be insufficient to examine only 19 journals, as Dedrick et al. (2009) and Luo et al. (2021) did in their review. Thus, I broadened the search scope and focused on four data sources related to education and psychology: APA PsyArticles, APA PsycInfo, Education Source, and Education Resources Information Center (ERIC). The keywords included "cross-classified," "cross-classified multilevel model," or "cross-classified random." The screening procedure is described below.

First, I identified studies using the noted keyword search for the four data sources. I removed any duplicate studies and reported the number of remaining studies. Next, I screened the title and abstract of the studies. Based on the inclusion and exclusion criteria, I filtered out studies that used CCREM only for methodological evaluations without demonstration application, used the cross-classified SEM, and used CCREM for meta-analysis or IRT models. If the title and abstract were insufficient to assess their suitability, I conducted a full-text review. Finally, all search and screening processes were presented using the PRISMA flow diagram using the `PRISMA_flowdiagram()` in `PRISMA2020` package (Haddaway et al., 2022; Page et al., 2021).

3.1.3 Coding Procedure

I coded articles using the following general dimensions: (1) study characteristics, (2) model assumptions, (3) model specifications, and (4) computational issues. The complete coding manual is listed in Table 3.1. The characteristics listed in the coding manual were selected based on the previous HLM systematic reviews and adjusted for the CCREM (Antonakis et al., 2021; Dedrick et al., 2009; Luo et al., 2021).

Table 3.1

Coding Manual

Characteristics	
(1) Study Characteristics	
Data type	Cross-sectional data, Longitudinal data
Number of Level-1 units per cluster	Numeric
Number of Level-2 clusters per dimension	Numeric
Data structure	Two-level, Three-level, Four-level
Model structure	Two-level, Three-level, Four-level
Clustering data structure	Two-way, Three-way, Four-way or more
Clustering model structure	Two-way, Three-way, Four-way or more
Rationale provided for using CCREM	Yes, No
Type of the outcome variable	Continuous, Binary, Ordinal
(2) Data consideration	
Exogeneity assumption	Assumption mentioned
	Assumption tested
	Response to the violation
	Consequences considered
	Corrective action taken
	Not discussed
	Not applicable
Homoscedasticity assumption	Assumption mentioned
	Assumption tested
	Response to the violation
	Consequences considered
	Corrective action taken
	Not discussed
	Not applicable
Normality assumption	Assumption mentioned
	Assumption tested
	Response to the violation

Table 3.1

(continued)

Characteristics	
	Consequences considered
	Corrective action taken
	Not discussed
	Not applicable
(3) Model Specification	
Centering at lower levels	Not centering
	Grand-mean Centering
	Cluster-mean Centering
	Not discussed
	Not applicable
Centering at the highest levels	Not centering
	Grand-mean Centering
	Not discussed
	Not applicable
Interaction between covariates examined	Level-1
	Level-2
	Cross-level
	No interaction
	Not applicable
Random interaction effects examined	Random interaction effects examined
	No random interaction effects
Random slope examined	Random slope examined
	No random slope
	Not applicable
(4) Computational issues	
Software	HLM
	MLwiN
	M*plus*
	R
	SAS
	SPSS

Table 3.1

(continued)

Characteristics	
	Stata
	Other
	Not reported
Estimation Method	ML
	REML
	Bayesian approach
	Other
	Not reported

Study Characteristics

The study characteristic factors focused on the cross-classified data structure, including whether the data is cross-sectional or longitudinal, the number of clusters for each dimension, and the number of level-1 units per cluster. I also recorded the levels of the data structure and the level of the actual model applied separately for instances where researchers might have reflected the real-world data structure in their models differently. For example, if the data structure consists of three levels, such as students nested within classrooms and classrooms nested within schools, researchers might model the data using a simplified two-level structure and exclude the school or the classroom level.

I further reported the number of cross-classified dimensions. The cluster dimensions represent the dimensions that are cross-classified at the same level. In cases where the clustering structure involves three or more dimensions, researchers might have omitted one or two dimensions which are assumed to have a negligible impact, resulting in a two-way clustering model. To account for this, I recorded the observed clustering data structure and the clustering model structure implemented by the researcher separately. Finally, I coded whether the rationale for employing the CCREM has been stated and the scale of the outcome variables.

Model Assumptions

In the model assumptions section, I coded whether the studies tested the CCREM assumptions, including exogeneity, homoscedasticity, and normality. Studies that mention the assumptions still might not test the assumptions. I thus specified whether the researchers actually tested the assumptions and how. Moreover, I detailed how the studies responded to potential violations of the assumptions.

Model Specification

In the model specification, I coded whether centering was used with each covariate at each level and captured the type of centering used in a study. Some studies might have employed standardization of the covariates in the model. In that case, I considered standardization a form of grand-mean centering (Hox et al., 2017). I also coded which fixed effects parameters were modeled as random (intercept and/or slopes) and whether random interaction effects were examined in the CCREM that was tested.

Computational Issues

Regarding computational issues, I reported which software the researchers used because defaults for estimation may vary depending on the software used. Also, I coded the type of estimation method used. The author conducted all coding.

3.2 Results

The systematic review collected CCREM studies published over ten years and explored the characteristics of CCREM studies. Figure 3.1 illustrates the PRISMA flowchart of the systematic review process. In the initial search from the data sources, I first identified a total of 337 CCREM studies. After removing 25 duplicates, 314 studies were identified. Next, I conducted the abstract screening and excluded three non-English studies, leaving 311 studies. These 311 studies were all eligible for inclusion and assessed via full-text screening. Of these, 93 studies were excluded for not meeting the inclusion criteria. The details of the exclusion are illustrated in Figure 3.1. For example, studies not specifically focused on CCREM, studies that aimed to provide instruction on CCREM methodology, and simulation studies without empirical data illustration were excluded. The final 218 studies were included in the systematic review.

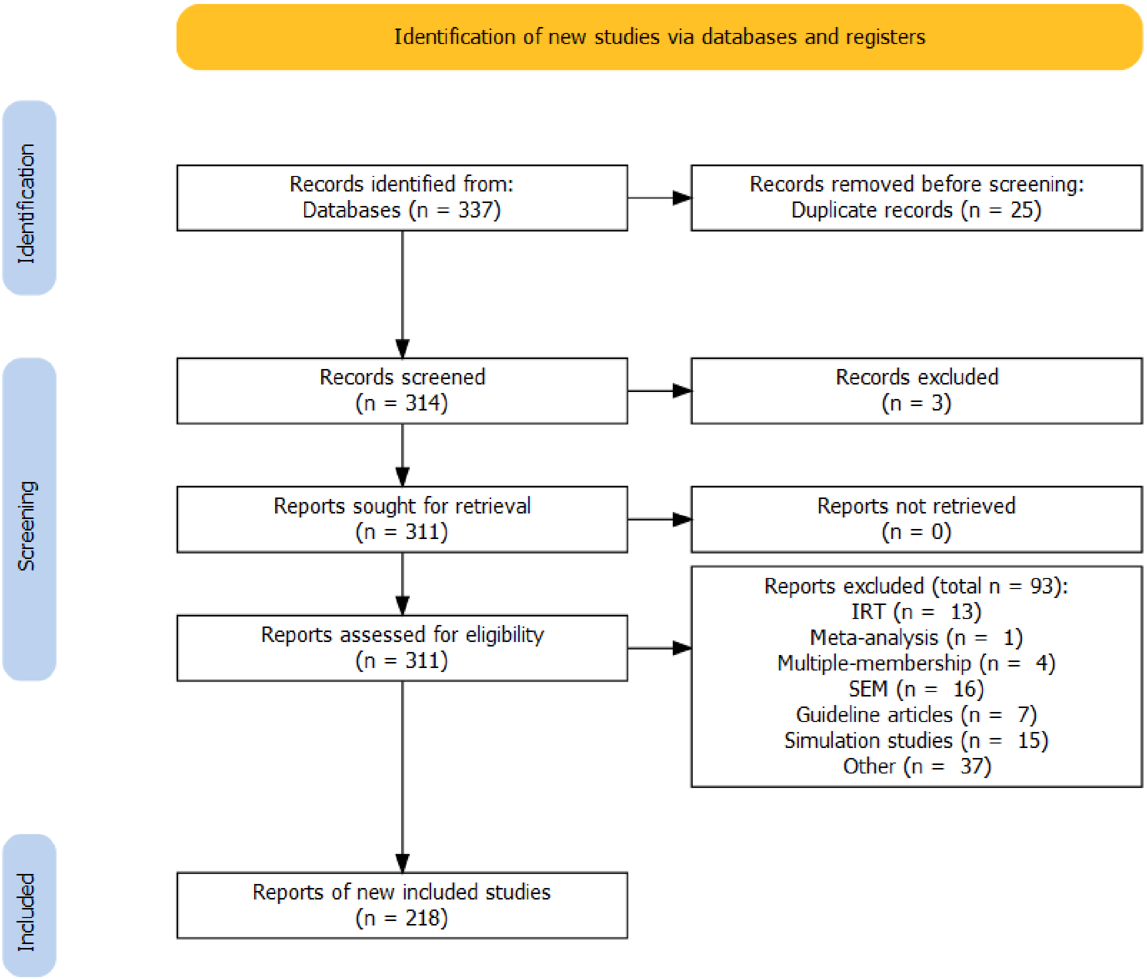
Identification of new studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 337)
Records removed before screening:
Duplicate records (n = 25)
Records screened
(n = 314)
Records excluded
(n = 3)
Reports sought for retrieval
(n = 311)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 311)
Reports excluded (total n = 93):
IRT (n = 13)
Meta-analysis (n = 1)
Multiple-membership (n = 4)
SEM (n = 16)
Guideline articles (n = 7)
Simulation studies (n = 15)
Other (n = 37)
Reports of new included studies
(n = 218)

Figure 3.1. PRISMA Flow Diagram Example

The selected studies were published in a total of 140 journals. Social Science and Medicine published the most CCREM studies ($N = 9$), followed by Health and

Place (N = 6), Journal of Adolescent Health and Social Indicators Research (N = 5), and Contemporary Educational Psychology, Frontiers in Psychology, Journal of Educational Psychology, Teachers College Record, and The International Journal of Behavioral Nutrition and Physical Activity (N = 4, each). I also plotted the number of CCREM studies published each year in Figure 3.2. The distribution of 218 studies shows that CCREM studies have been published consistently over the past decade, with an average of 21.8 studies per year.

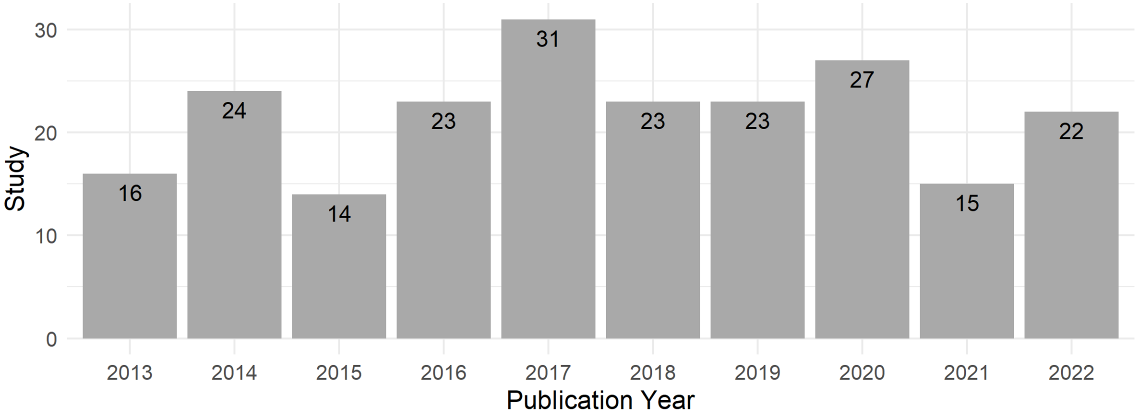

Figure 3.2. Study Distribution by Publication Year

3.2.1 Study Characteristics

Table 3.2 shows study characteristics for the included studies. Among the selected studies, cross-sectional data (N = 167, 77%) were more common than longitudinal data (N = 51, 23%). The example of cross-sectional data included cases where individuals were nested in different types of clustering dimensions (e.g., school, region, or rater) or where observed measures were cross-classified into different types of higher-level clusters (e.g., individual, item, or region).

I also investigated the size of the cross-classified datasets by examining the number of level-1 units per cluster and the number of clusters in each dimension. To account for outliers and the non-normal distribution of the study size, I calculated the first quartile (Q1), second quartile (Q2), and third quartile (Q3) for each level and

Table 3.2

Study Characteristics

Characteristics			
Data type:			
Cross-sectional data	167 (77%)		
Longitudinal data	51 (23%)		
	Q1	Q2	Q3
Number of level-1 units per cluster	18	95	436
Number of level-2 clusters per dimension			
Smaller cluster size	21	52	158
Larger cluster size	80	303	1,418
	Two-level	Three-level	Four-level
Multilevel Data structure	185 (85%)	23 (11%)	10 (5%)
Multilevel Model structure	187 (86%)	24 (11%)	7 (3%)
	Two-way	Three-way	Multi-way
Clustering data structure	176 (81%)	37 (17%)	5 (2%)
Clustering model structure	190 (87%)	23 (11%)	5 (2%)
Type of the outcome variable			
Continuous	150 (69%)		
Binary	66 (30%)		
Ordinal	7 (3%)		
The rationale provided for using CCREM			
Yes	218		
No	0		

dimension (see Table 3.2). Regarding the number of level-1 units per cluster in the smaller dimension, the interquartile range (Q3 - Q1) was 418. The range of level-1 units per cluster spanned from 2 to 4,022,312, with an average of 25,749.

The number of clusters per dimension was collected as the smaller dimension and larger dimension separately, considering the unequal size of each dimension. In the smaller clustering dimension, the interquartile range was 137, with the minimum and maximum numbers being 2 and 12,686, respectively. The average number of level-2 clusters in this smaller dimension was 448. In the larger clustering dimension, the interquartile range was 1,338. The minimum and maximum were 6 and 636,202, respectively, with an average of 7,764.98.

The results indicate that the number of clusters varies widely, which may be due to the diverse types of clusters. For example, Kendler et al. (2015) used data that incorporated ten distinct Swedish nationwide registries and healthcare data using unique individual IDs. The study focused on individuals born in Sweden between 1975 and 1990 who were residing in Sweden at the end of 1990, and the individuals were nested within communities and households. This study, with its large national data set, reported the largest number of clusters among the selected studies, which was 636,202 households.

And a common type of cluster dimension was often the individual. For example, in Patchan et al. (2016), the data consisted of peer feedback on students' papers, which were nested within the reviewers and authors. In Thornton III et al. (2019), the dataset included assessments on promotion exams, which were nested within the assessees and assessors. Weiser (2013) examined individuals with reading disabilities who were nested within the intervention teachers and classroom teachers.

I recorded the number of levels in each study's cross-classified data structure and noted the corresponding levels specified in the model. Two-level data was observed most frequently, accounting for 85% ($N = 185$) of the studies. Three-level data followed with 11% ($N = 23$) of the data structures, and four-level data accounted for 5% ($N = 10$). To illustrate, Gilbert et al. (2016) had a four-level data structure where time points (level-1) were nested within students (level-2) who were cross-classified in second-grade and third-grade classrooms (level-3), which were again nested within schools (level-4).

In some cases, when certain levels exhibit low variance (i.e., low ICC values), researchers may choose to exclude those levels from their data structure and specify modified versions of CCREM. To account for this, I provided separate reporting of the model structure levels. Among data structures with two and three levels, no studies excluded any levels. However, in the case of four-level data structures, three studies employed CCREM with fewer levels than the original four-level. For instance, Dronkers et al. (2014) analyzed four-level data using a three-level CCREM. Also, Cafri et al. (2015) and Groenewegen et al. (2018) utilized a two-level CCREM on four-level cross-classified data. As a result, the model structure results revealed 187 two-level CCREMs (86%) and 24 three-level CCREMs (11%).

Further, the number of cluster dimensions in the analyzed data was reported. Among the included studies, the majority of studies had two-way cluster dimensions (81%, $N = 176$), followed by three-way (17%, $N = 37$) and four-way dimensions (2%, $N = 5$). However, not all cluster dimensions were modeled in the analyzed data. Of studies with three-way clustering dimensions, for example, 14 studies decided to omit one of the clustering dimensions and ended up with two-way dimensional CCREM.

Specifically, the age-period-cohort (APC) analysis was a common type of three-way cluster dimension where individuals are nested within age, period, and cohort. However, considering the strong dependence between age and cohort with a period, many studies analyzed these data using a two-way model (e.g., Attell, 2020; Beck et al., 2014; Hayward & Krause, 2015; Lin et al., 2014; Zhang, 2017), as recommended by Yang and Land (2008). For four-way clustering dimensions and above, there was no case for excluding clustering dimensions. As a result, the number of two-way dimensional CCREMs was 190 (87%), while the number of three-way CCREMs was 23 (11%). The number of four-way or above (i.e., multiway) remained at 4 studies (2%)

In terms of the type of dependent variable, the majority used in CCREM research were continuous variables ($N = 150$), which accounted for 69% of the total number of studies. The remaining studies utilized categorical outcome variables, with 66 studies (30%) using dichotomous variables and seven studies (3%) employing ordinal variables. For the analysis of these categorical variables, corresponding methods such as the cross-classified multilevel logistic model or the cross-classified multilevel ordered logit model were employed instead of linear CCREMs. Finally, all studies clearly stated their reasons for using the CCREM (218, 100%).

3.2.2 Model Assumptions

Table 3.3 presents the extent to which CCREM assumptions were considered in the studies analyzed. Of the three assumptions examined (exogeneity, heterogeneity, and normality), the exogeneity assumption was the least mentioned ($N = 9$, 4%) and never tested ($N = 0$). Only a few studies considered or corrected their results due to potential violation of the exogeneity assumption ($N = 3$, less than 2%).

For example, Silber et al. (2021) was the only study that considered the consequence of the violation, highlighting the possibility of endogeneity in certain variables and conducting a sensitivity analysis to evaluate the robustness of their results. Two other studies implemented corrections for the endogeneity issue: Nisic and Melzer (2016) used the Mundlak model, a variation of the CRE model, and Fleischmann et al. (2022) used cluster-mean centering (within-RE model). However, most studies did not discuss the exogeneity assumption ($N = 202$, 93%). Three studies were not applicable for evaluating the exogeneity assumption because they did not include any variables in the model ($N = 4$, 2%). These studies only examined the variance of the random effects.

Table 3.3

Model Assumptions

Characteristics			
	Exogeneity	Homoscedasticity	Normality
Assumption mentioned	9 (4%)	49 (22%)	58 (27%)
Assumption tested	0 (0%)	9 (4%)	18 (8%)
Response to the violation:			
Consequences considered	1 (0.5%)	0 (0%)	0 (0%)
Corrective action taken	2 (1%)	1 (0.5%)	10 (5%)
Not discussed	202 (93%)	159 (73%)	132 (61%)
Not applicable	4 (2%)	0 (0%)	0 (0%)

The homoscedasticity and normality assumptions were mentioned more frequently than the exogeneity assumption in CCREM studies. Two assumptions are often mentioned when describing a level-2 random effect or level-1 residual, typically stating that the random effects and residuals are assumed to follow a normal distribution with a mean and variance of a certain value (e.g., Castellaneta & Gottschalg, 2016; Lei et al., 2018). Specifically, the homoscedasticity and the normality assumptions were reported in 49 studies (22%) and 58 studies (27%), respectively. In addition, 9 studies (4%) tested for heterogeneity, and 18 studies (8%) assessed normality. These assumptions were evaluated using various methods, such as residual plots or Q-Q plots at each level (e.g., Dunn et al., 2015; Goodale et al., 2019; van Berkel et al., 2022). Bayer-Oglesby et al. (2022) used the modified Breusch-Pagan test (Abdul-Hameed & Matanmi, 2021; Breusch & Pagan, 1979) to test heterogeneity of residual variance. Although some studies did not specify the test method, I considered the assumption to be tested when the authors mentioned that it was tested (e.g., Kim & Sax, 2014).

Violation of the normality assumption was handled more often ($N = 10$, 5%) than the homoscedasticity assumption ($N = 1$, 0.5%) when there was a violation of these assumptions. The violation of the homoscedasticity assumption was corrected only in Patton (2019) by excluding existing outliers. The normality assumption was corrected in several studies by removing outliers (e.g., Baird et al., 2017; Bauer et al., 2021; D'Haese et al., 2014; Thrash et al., 2017) and using a log transformation for the outcome (e.g., Morton et al., 2016; Patton, 2019). One study employed sensitivity analyses to address normality violations (e.g., Allensworth & Luppescu, 2018). However, a considerable number of studies did not report on these assumptions. The heterogeneity assumption was not discussed in 159 studies (73%), and the normality assumption was not discussed in 132 studies (61%).

3.2.3 Model Specification

In the model specification (see Table 3.4), I examined how studies using the CCREM specified their models in terms of centering methods, interaction effects between covariates, random interaction effects, and random slope. I first explored how centering methods were included in the model at the lowest level, level-1, and at higher levels depending on the levels included in the model. Not centering category indicates that the authors considered centering but ultimately did not center covariates in their models. The number of cases when covariates were not centered was small for every level. Six studies (3%) used the raw predictor at level-1, and 13 studies (6%) used the raw predictor without centering at higher levels.

Grand-mean centering was the most commonly used method regardless of the level at which the covariate was included. At level-1, grand-mean centering was used in 54 studies (25%), of which 34 studies effectively used grand-mean centering through the use of a standardized variable. Grand-mean centering was more commonly used at higher levels than at level-1, with 65 studies (30%) utilizing this method. Of those, 23 studies used standardization as grand-mean centering.

Table 3.4

Model Specification

Characteristics		
Centering at	Level-1	Highest level
Not centering	6 (3%)	12 (6%)
Grand-mean Centering	54 (25%)	65 (30%)
Cluster-mean Centering	19 (9%)	NA
Not discussed	117 (54%)	125 (57%)
Not applicable	26 (12%)	16 (7%)
Interaction between covariates examined		
Level-1	22 (10%)	
Level-2 or higher	49 (22%)	
Cross-level	55 (25%)	
No interaction	104 (48%)	
Not applicable	4 (2%)	
Random interaction effects examined		
Random interaction effects examined	11 (5%)	
No random interaction effects	205 (94%)	
Random slope examined		
Random slope examined	41 (19%)	
No random slope	175 (80%)	
Not applicable	2 (1%)	

At level-1, cluster-mean centering was used in 19 studies (9%). Cluster-mean centering was applied in various ways, including centering around a specific time point (e.g., Dağli & Jones, 2013; Francis et al., 2018) or using the cluster-median instead of the cluster-mean (e.g., Beck et al., 2014; Lin et al., 2016). Some studies also standardized covariates by cluster (e.g., Evans & Fite, 2019) or presented results using both cluster-mean and grand-mean centering (e.g., Fleischmann et al., 2022; Vagi et al., 2017).

It is unexpected that the cluster-mean centering in these studies was not implemented by centering on multiple dimensions. These studies calculated the cluster-mean-centered covariate based on one dimension, and the remaining clustering dimension was not accounted for in the cluster-mean centering. Pedersen et al. (2018) and Sharp et al. (2015) were the only studies that considered multiple clustering dimensions by using the CRE model as the cluster-mean centering. These two studies conducted the CRE model by including multiple cluster means in the model while using the raw covariate. Cluster-mean centering is not applicable at the highest level, and thus it was only reported in lower levels that nested within a cluster.

More than half of the studies did not consider any centering method. Specifically, 117 studies (54%) did not mention any centering at level-1, and 125 studies (57%), which had the option to use grand-mean centering at level-2 or higher, also did not mention centering. 26 studies at level-1 (12%) and 16 studies at higher levels (7%) were classified as "Not applicable" because they had no covariates included in the higher level of the model.

Regarding the estimation of interaction effects, approximately half of the studies included interaction terms between variables. Specifically, 22 studies (10%) estimated interaction terms at level-1, while 49 studies (22%) estimated interaction terms between covariates of different dimensions at level-2 or higher. 54 studies (25%) examined cross-level interaction terms. However, 104 (48%) studies did not use any interaction term.

In addition to interaction effects between fixed effects, it is also possible to include interaction terms between cross-classified dimensions' random effects (i.e., random interaction effects). Only 11 studies (5%) specified a random interaction term. Two of these studies included a random interaction term to measure dyadic variance between individuals (e.g., Kim et al., 2015; van Braak et al., 2021). The remaining studies mentioned that the interaction was included for estimating the interaction between two clustering dimensions, such as teachers and students (e.g., Feistauer & Richter, 2017, 2018) or perceiver and target (e.g., Hehman & Sutherland, 2017; Xie et al., 2019). However, random interaction effects are still uncommon, and most studies ($N = 205$, 94%) did not discuss nor model them.

Lastly, I investigated whether the studies included a random slope in their models. Of the total, 41 studies (19%) included random slopes, allowing the slope to vary across clusters. However, the majority of the studies ($N = 175$, 80%) utilized the random intercept model instead. Two studies with unconditional CCREM were reported as not applicable (1%).

3.2.4 Computational Issues

In the computational issues section, I examined the software and estimation methods in selected studies (see Table 3.5). If multiple software programs were mentioned and the researcher did not specify which software was used in the CCREM analysis, all software was recorded. Thus, the sum of the number of software programs used may exceed the number of studies. The results indicated a diverse range of software employed for CCREM analyses. In the selected studies, `MLwiN` and `R` account for approximately 40% of the software used. Specifically, 47 studies (22%) used `MLwiN`, with 11 studies using `MLwiN` via `Stata` and one study using `MLwiN` via `R`.

`R` was used in 42 studies (19%), with 31 studies using the `lmer4` package (Bates et al., 2015), two studies with the `MCMCglmm` package (Hadfield, 2010) and one study using the `Stan` package (Stan Development Team, 2023). Two studies classified

Table 3.5

Computational Issues

Characteristics		
Software		
HLM	14	(6%)
MLwiN	47	(22%)
M*plus*	7	(3%)
R	42	(19%)
SAS	22	(10%)
SPSS	10	(5%)
Stata	24	(11%)
Other	2	(1%)
Not reported	55	(25%)
Estimation Method		
ML	22	(10%)
REML	20	(9%)
Bayesian approach	55	(25%)
Other	1	(0.5%)
Not reported	120	(55%)

as "Other" used Stat-JR and WinBUGS. However, a significant proportion of studies (55, 25%) did not report which software they used.

To explore possible trends in software usage, I examined the frequency of software usage by year. Table 3.6 displays the most commonly used software, listed in order of frequency. Although the trend is unclear, there has been a slight decline in the use of MLwiN throughout the 2020s, while the number of papers using R has slightly increased. Stata and SAS showed a consistent level of use over time.

Finally, I examined the estimation methods used in CCREM studies. The three most common estimation methods used were Maximum Likelihood (ML; $N = 22$, 10%), Restricted Maximum Likelihood (REML; $N = 20$, 9%), and Bayesian approaches, with the latter being the most widely used ($N = 54$, 25%). The choice of estimation method may be related to the software utilized in the study. For instance, 39 out of the 47 studies that used MLwiN as their software reported adopting a Bayesian approach using Markov Chain Monte Carlo (MCMC) estimation procedures.

Table 3.6

Frequency of Software Usage by Year

Software	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	Total
MLwiN	3	8	4	7	7	3	7	4	1	3	47
R	2	2	1	4	4	7	4	8	2	8	42
Stata	1	5	4	3	1	4	1	1	3	1	24
SAS	3	1	2	2	3	3	1	3	2	2	22
HLM	3	2	0	3	1	0	3	1	0	1	14
SPSS	0	1	1	0	2	1	1	0	2	2	10
Mplus	0	0	0	0	1	1	2	2	0	1	7

However, the most common estimation method in other software packages was not apparent. A single study that reported using penalized quasi-likelihood (PQL) for handling binary outcomes was classified as "Other." However, over half of the studies ($N = 120$, 55%) did not report the estimation method.

These systematic review results highlighted the methodological characteristics of CCREM in various research fields and shed light on the assumptions and centering use in CCREM analyses. However, little attention has been given to the exogeneity assumption underlying CCREM, which is critical for obtaining unbiased and consistent estimates. Further, the cluster-mean centering approach to address endogeneity seems not implemented correctly in CCREM. Therefore, in the next section, I describe the simulation study results conducted to compare alternative modeling approaches that can relax the exogeneity assumption and provide more robust estimates in the presence of endogeneity.

Chapter 4

Simulation Study

4.1 Methods

In the Monte Carlo simulation study, I compared the estimation of pooled, within- and between-cluster effects of level-1 covariates for alternatives designed to handle endogeneity using various centering methods. I examined nine approaches for the model and level-1 covariate centering paired with estimating a CCREM include: (1) not centering, (2) grand-mean centering, (3) the within-RE model, (4) the CRE model, (5) the within-cell RE model, (6) the correlated-cell RE model, (7) the two-way FE-CRVE with an uncentered covariate, (8) the FE-RE hybrid approach using the school dimension as the fixed effect and the neighborhood dimension as the random effect, and (9) the FE-RE hybrid approach using the school dimension as the random effect and the neighborhood dimension as the fixed effect. Both hybrid models used an uncentered covariate.

I specifically focused on cross-sectional cross-classified data assuming an unbalanced design, in which the number of units nested within a cluster often differs within each cross-classified dimension. A balanced design, such as panel data, may produce more optimal results than the analysis of unbalanced data. However, I aimed to derive more generalizable and practical conclusions by generating unbalanced, more challenging, realistic data. The CCREMs used to generate the data and to estimate the models were random intercept models.

The primary focus of the simulation study was on the performance of within-cluster effect coefficients for level-1 covariates estimated using alternative approaches (3) to (9). For the model with the (1) uncentered covariate and (2) grand-mean-centered covariate, the performance of the pooled effect of the level-1 covariate was evaluated. The between-cluster effects were estimated only using (4) the CRE and

(6) the correlated-cell RE models.

The performance of cell-interaction effects estimated only from the correlated-cell RE model was not evaluated. Including the cell means in the model serves a more important role as additional covariates rather than being a primary research interest. The level-2 covariate was also not a focus and was not included in this simulation study. In two-level CCREM, only not centering or grand-mean centering can be used for level-2 covariates; cluster-mean centering and cell-mean centering cannot be employed.

4.1.1 Data Generation

Based on the characteristics of the empirical data, I generated two-level cross-classified data mimicking students cross-classified by schools and neighborhoods. I first generated student IDs ($i = 1, 2, ..., n_j$) for each school ($j = 1, 2, ..., J$). Then, I assigned students from each school to the neighborhoods according to the following equation,

$$k_{ij} = j(\frac{K}{J}) + x, \tag{4.1}$$

where k_{ij} is the neighborhood ID that student i attending school j is assigned. I used $\frac{K}{J}$, the ratio of the total number of schools J and neighborhoods K, so that students in a school can be distributed to the neighborhood IDs proportional to the school ID value. The term x represents a random number generated from a uniform distribution of the interval [1, sparsity × K], allowing students in one school to be distributed evenly in a few neighborhoods, not concentrated in one neighborhood.

According to Raudenbush and Bryk (2002) and Paterson (1991)'s cross-classified data structure, I adopted the degree of 10% as the sparsity used for data generation in my study. For example, when the number of schools and neighborhoods are 30 and 60, respectively, students in school ID $=$ 1 are randomly assigned to neighborhood IDs between 3 and 8 (c.f., $k_{ij} = 1(\frac{60}{30}) + x$, where $x \sim [1, 0.1 \times 60]$), and students in school ID $=$ 2 are randomly assigned to neighborhood IDs between 5 and 10 (c.f.,

$k_{ij} = 2(\frac{60}{30}) + x$, where $x \sim [1, 0.1 \times 60]$).[1]

Next, I generated values for the level-1 covariate. In cross-classified data, different sources affect the level-1 covariate: within-cluster effects, two between-cluster effects, and the interaction effects between clusters from each dimension. Using the empirical examples, I examined the variability of each effect using a random intercept model with the predictor variable as an outcome (see footnotes on Tables 2.2 and 2.4), and found that the IUCC of the interaction effects was not always smaller than the IUCC of the clustering dimensions. In order to reflect these variabilities within the level-1 covariate, I arbitrarily selected the generating values so that the variances for sources underlying values on the covariate totals 100: i.e., the within-cluster variability $X_w \sim N(0, 25)$, each of the between-cluster variances $X_{b,j}$ and $X_{b,k} \sim N(0, 25)$, respectively, and the cell-interaction effects' variability $X_{j \times k} \sim N(0, 25)$. I manipulated the proportion of variance assigned to the clustering dimensions and their interaction in the covariate equal to the variance in within-cluster variability to simplify the proportion of each effect. Finally, I summed the effects from four sources to generate values on the level-1 covariate.

Then, I generated level-2 random effects, including random effects for the school and neighborhood dimensions, random interaction effects between school and neighborhood dimensions, and level-1 errors. Under the scenarios where the exogeneity assumption was met, all random effects were independently sampled from a normal distribution and uncorrelated with covariates. In other words, random school and neighborhood effects were sampled from independent normal distributions with means of 0 and variances of τ_{j00} and τ_{k00}, respectively. The random interaction effect was generated to follow a normal distribution with a mean of 0 and a variance a $\tau_{(jk)00}$. The level-1 errors were sampled from a normal distribution with a mean of 0 and a variance of σ^2.

[1] If the school ID is 30, the potential neighborhood IDs range from 61 to 66, which exceeds neighborhood ID = 60. Therefore, if the generated neighborhood ID is greater than $K = 60$, I subtracted 60 from the ID value so that the possible neighborhood ID range is between 1 and 6.

Under conditions in which the exogeneity assumption is violated, I generated the random neighborhood effect c_{00k} to be correlated with the neighborhood source of values for the level 1 covariate $X_{b,k}$:

$$c_{00k} = r \times \sqrt{\frac{\tau_{k00}}{25}} \times X_{b,k} + \upsilon_{00k}, \tag{4.2}$$

where r is the correlation between the random neighborhood effects and the neighborhood components of the level-1 covariates, and υ_{00k} follows a normal distribution with a mean of zero and variance of $(1-r^2)\tau_{k00}$. The endogeneity may also be present in one or both cross-classified factors or interaction terms. However, in this simulation study, endogeneity was introduced only into the neighborhood dimension as a starting point for this line of research.

After generating all the components, I calculated the outcome variable $Y_{i(jk)}$ based on the correlated-cell RE model with a random intercept and a level-1 covariate. In a scalar form, the correlated-cell RE model in Equation 2.51 can be represented by

$$Y_{i(jk)} = \gamma_w X_w + \gamma_{b,j}\bar{X}_{b,j} + \gamma_{b,k}\bar{X}_{b,k} + \gamma_{b,jk}\bar{X}_{jk} + b_{0j0} + c_{00k} + d_{0jk} + e_{i(jk)}, \tag{4.3}$$

with the fixed effect intercept value generated to be zero. The correlated-cell RE model was chosen so that the within-cluster, between-cluster, and cell-interaction effects of the covariate can be generated, respectively.

4.1.2 Conditions

Table 4.1 lists the experimental factors manipulated in the simulation study. Taking into account the data structure in the empirical example, the systematic review, and conditions employed in related previous CCREM-focused simulation studies, I selected experimental factors anticipated to influence the performance of the models and centering methods. The experimental factors included: (1) CCREM assumptions, (2) coefficient size for within-cluster, between-cluster, and cell-interaction effects, (3) the number of clusters per dimension, (4) the number of level-1 students nested within level-2 schools, (5) the IUCC for the between neighborhoods and cell-interaction effects.

Table 4.1

Simulation Conditions

Experimental Factors	Levels
CCREM assumptions	All assumptions met Exogeneity assumption violated
Coefficients size	.01 (small), .02 (medium), or .04 (large)
Number of level-2 clusters, schools J × neighborhoods K	20×70 (small), 70×245 (medium), or 150×525 (large)
Number of level-1 students per school n_j	30 (small) or 100 (large)
Neighborhood IUCC	.05 (small), .15 (medium), or .25 (large)
Cell-interaction IUCC	.00 (small), .05 (medium), or .15 (large)

Note. The ratio between level-2 clustering dimensions is 3.5:1. School IUCC is set to .05. The total number of conditions is $2 \times 3 \times 3 \times 2 \times 3 \times 3 = 324$.

CCREM Assumptions

The first experimental factor is whether exogeneity can be assumed in the CCREM. I generated conditions where the exogeneity assumption is met along with other CCREM assumptions and other conditions in which the exogeneity assumption is violated. For scenarios in which the exogeneity assumption is not violated, I generated a random intercept CCREM in which the random effects are independent of the covariate in the model. For conditions in which exogeneity is violated, data was generated to fit an identical CCREM, except that one of the dimension's (the neighborhood's) random effects was sampled as correlated with the level-1 covariate (see Equation 4.2).

In the empirical data of Raudenbush and Bryk (2002), the correlation between the covariates and the random neighborhood effects was between 0.056 and 0.078. In Paterson (1991) data, the correlation was between 0.143 and 0.165. A larger correlation value was investigated in previous simulation studies focused on the exogeneity assumption in multilevel models. For example, Castellano et al. (2014) used correlation values of 0.2, 0.4, and 0.6 when they examined alternatives for handling endogeneity in hierarchical data analyses. Their condition values were based on the correlation between the school random effects and the socioeconomic status found in the 1982 High School and Beyond survey data, which was 0.36. Thus, considering these previous studies, I used a value of 0.4 to generate the correlation between the random effects and the covariates.

Coefficient Size

In Lee and Pustejovsky (2023), the impact of the value of the coefficient for the level-1 covariate on the model's performance was negligible. However, the coefficients examined in the previous study were the pooled effects of the covariates. The value of each effect composing the pooled effect, i.e., within-cluster effects, between-cluster effects, and cell-interaction effects, could still impact the model's performance. Thus, I manipulated the coefficient size and confirmed that the corresponding correlations are reasonable. The value of the within-cluster effect can be calculated as $ES = \gamma_w \times \frac{sd(X_{i(jk)})}{sd(Y_{i(jk)})}$. For example, when the target coefficient value is 0.01, the within-cluster effect size would be $\gamma_w \times \frac{sd(X_{i(jk)})}{sd(Y_{i(jk)})} = 0.01 \times \frac{10}{1} = 0.1$. This effect size represents the correlation between the within-cluster portion of the covariates and the corresponding level-1 error components.

However, the calculation of between-cluster effects is more complicated in CCREM. HLM calculates the between-cluster effect based on the within-cluster effect and the constraint of level-1 and level-2 correlation (Raudenbush & Bryk, 2002, Chapter 3). To my knowledge, however, no previous study has demonstrated how to calculate between-cluster effects in CCREM. Therefore, I examined the empirical effect size for between-cluster and cell-interaction effects through simulated data to determine whether these are plausible values under the manipulated conditions.

In the simulation study, I set the target coefficient value as 0.01, 0.02, and 0.04.

The within-cluster, between-cluster, and cell-interaction coefficients were generated as equal values to ensure stability in the coefficients. I fixed the number of clusters at 200 and set the number of replications to 10,000 to generate a sufficiently large population of cross-classified data. The empirical correlations between the covariates and the corresponding errors were calculated at each cluster and cell level. Figure 4.1 shows the results of the empirical correlation and its estimate. This graph includes only those simulation conditions that affected the correlation based on the ANOVA: Coefficient γ, Neighborhood IUCC, and Cell-Interaction IUCC. The empirical correlations were all less than one and had the same order of magnitude as the ES parameter. Thus, I concluded the effect size of 0.1, 0.2, and 0.4 would be plausible.

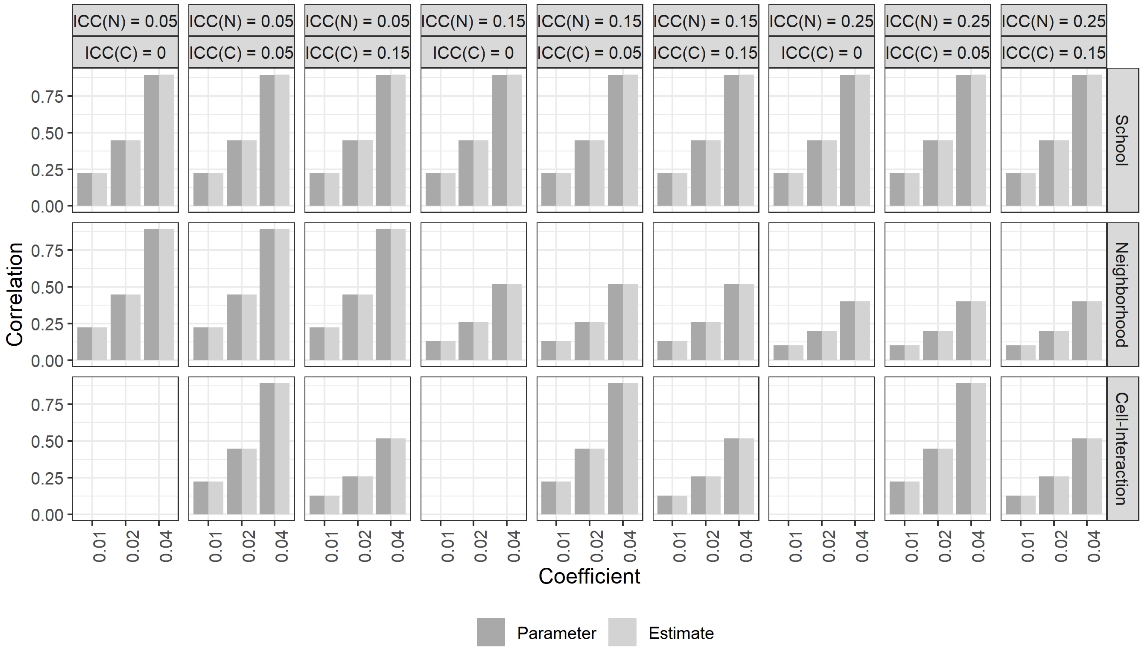

Figure 4.1. Correlation Parameter and Estimates for School, Neighborhood, and Cell-Interaction Dimension.
Note. In the columns, ICC(N) and ICC(C) denote the IUCC conditions for neighborhood and cell-interaction clustering dimensions, respectively. Columns 1, 4, and 7 of the third line are blank because the empirical effect size for the cell interaction effect was not calculated when the cell IUCC was zero.

Number of Level-2 Clusters

In an unbalanced cross-classified dataset, the number of clusters in each dimension often varies. For example, in Raudenbush and Bryk (2002)'s empirical example, the number of schools was 17, whereas the number of neighborhoods was 524. In Paterson (1991), the number of primary schools was 148, and the number of secondary schools was 19. Thus, in determining the number of schools, I referred to the dimension with the smaller number of clusters based on the data collected in the systematic review.

Using the results of my systematic review, I calculated the first, second, and third quantiles of the smaller number of clusters for the school dimensions (21, 52, and 158) and the larger number of clusters for the neighborhood dimensions (80, 303, and 1,418). Each quantile represents a small, medium, and large condition for the number of schools, respectively. Also, I considered the average ratio of 3.4 between the two clustering dimensions to determine the number of clusters for the dimension of the larger number of clusters proportional to the smaller number of clustering dimensions. Considering these values, I used the condition of the smaller cluster as 20, 70, and 150 and calculated the larger number of clusters (e.g., neighborhoods) by multiplying the number of schools by 3.5 (i.e., 70, 245, and 525) to set the conditions closest to the empirical data conditions.

Number of Level-1 Students per School

In the systematic review, the first, second, and third quantiles of level-1 units per cluster were 18, 95, and 436, respectively. On the other hand, the simulation condition of the number of level-1 units per cluster in prior simulation studies has typically been 20 and 40 (Meyers & Beretvas, 2006) or even as low as 10 (Luo & Kwok, 2009). Considering both empirical and simulation study values with an unbalanced cross-classified design, I used two values for the number of level-1 units per school: 30 and 100.

The number of individuals per cluster is likely unbalanced in real world data. In order to reflect the unequal cluster sizes in real data, I sampled the sample size per cluster from a normal distribution, with an average of the determined value and the standard deviation of the value multiplied by 0.3. For example, if the condition was 30, then I generated the actual sample size per cluster by sampling from the normal distribution with a mean of 30 and a standard deviation of $30 \times 0.3 = 9$. This allowed the sample size per cluster to vary, creating an unbalanced cluster size design.

Neighborhood and Cell-Interaction IUCC

Based on empirical applications and multilevel modeling textbooks, IUCC values for CCREM ranged from .01 to .24 (Meyers & Beretvas, 2006). Previous simulation studies by Beretvas and Murphy (2013) and Meyers and Beretvas (2006) investigated conditional IUCC values of 0.05, 0.15, or 0.3. Further, IUCC was often different per clustering dimension. In order to reflect the wide range of IUCC values in empirical studies and the unbalanced IUCC per dimension, I fixed a school IUCC generating value of 0.05 and varied only the neighborhood IUCC and cell-interaction IUCC values based on the condition. I used three conditional IUCC generating values of 0.05, 0.15, and 0.25 for the neighborhood dimension.

In the empirical applications in the previous section, the IUCC for the random interaction effect varied between 0.08 to 0.25 and was not always smaller than the school or neighborhood IUCC (see Tables 2.2 and 2.4). Therefore, for the IUCC condition generating values for the cell-interaction IUCC, I used 0.05 and 0.15 as in the neighborhood dimension and added a condition of zero to examine a case in which there is no cell-interaction effect.

Distribution of the Number of Students in Cells

The generated cells formed based on the above conditions, i.e., the combinations between the school and neighborhood clustering dimensions, should contain at least one level-1 student. An empty cell indicates that the corresponding combination was not generated correctly. In the data generation process, I examined the distribution of the number of students present in the generated cells, as shown in Figure 4.2 and Table 4.2.

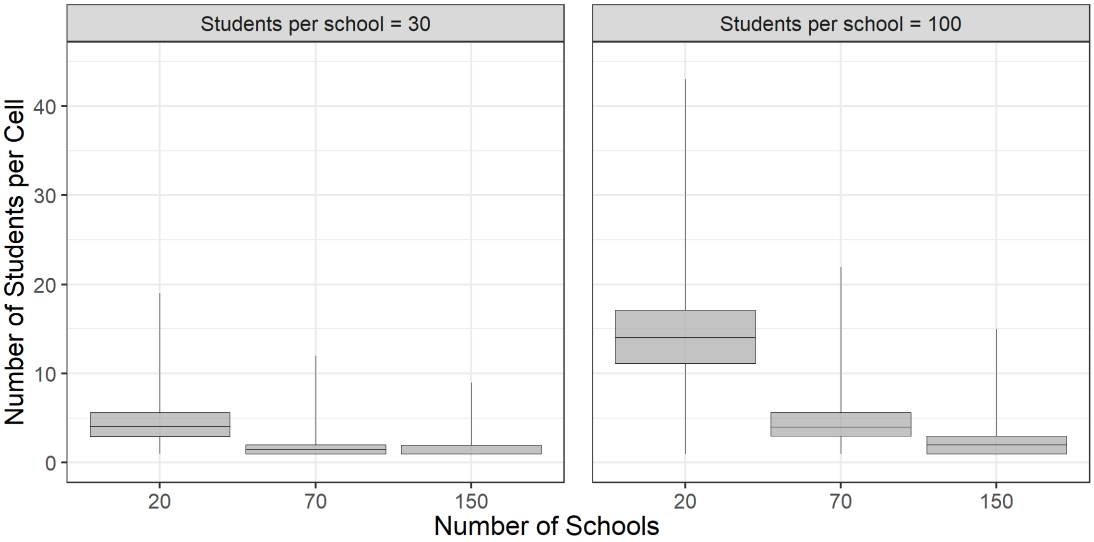

Figure 4.2. Distribution of the Number of Students in Cells by the Number of Schools and Number of Students per School

Table 4.2

Distribution of the Number of Students in Cells by the Number of Schools and Number of Students per School

School	Students/School	Min	Q1	Q2	Q3	Max	Mean	SD
20	30	1	2.95	4.07	5.65	19	4.35	2.04
20	100	1	11.2	14.0	17.1	43	14.3	4.47
70	30	1	1	1.47	2	12	1.76	0.939
70	100	1	2.98	4.00	5.62	22	4.25	2.10
150	30	1	1	1	1.94	9	1.32	0.588
150	100	1	1	2	3	15	2.28	1.27

Figure 4.2 and Table 4.2 illustrate that each cell contains at least one level-1 student. However, when the number of level-1 students was small and the number of

clusters was large, the number of students within the cell tended to be low as one. The presence of only one student in a cell can lead to convergence issues, particularly when employing cell-mean centering. In such cases, these cells with a single student may result in zero variance after subtracting the cell mean from the covariate value. I discussed the model convergence issue in more detail in the results section.

4.1.3 Performance Criteria

I evaluated the relative performance of CCREM centering methods, the two-way FE-CRVE model, and the FE-RE hybrid model with an uncentered covariate estimating level-1 coefficients using the following performance criteria: (1) relative and absolute parameter bias, (2) root mean square error (RMSE), and (3) relative bias of SE (Morris et al., 2019). First, relative parameter bias measures the relative difference in the value of an estimate (T) relative to the true value (θ),

$$\frac{\bar{T}-\theta}{\theta}, \tag{4.4}$$

where the acceptable cut-off value was suggested to be ±0.05 (Hoogland & Boomsma, 1998).

Next, absolute parameter bias was calculated to compare the bias of each model directly. Contrary to the relative parameter bias that calculates the proportion of the estimate from the true value, the absolute parameter bias calculates the difference between the two values,

$$\bar{T}-\theta. \tag{4.5}$$

Absolute parameter bias does not have a suggested criterion. However, a model with smaller values is regarded as having better performance.

The RMSE of the coefficient estimator describes the overall accuracy of estimation that accounts for both bias and the variance of the actual parameter estimates:

$$\sqrt{\frac{1}{n_{sim}}\sum_{i=1}^{n_{sim}}(T_i-\theta)^2}, \tag{4.6}$$

where n_{sim} is the number of replications. RMSE can describe the SE of the simulation estimates at the same scale as the parameter estimates. Smaller RMSE values represent greater accuracy, i.e., less bias and better precision (Burton et al., 2006; Morris et al., 2019).

Finally, the relative bias of SE quantifies the difference in SE estimates from the true (empirical) standard error relative to the scale of the true parameter,

$$\frac{\bar{SE_T}-S_T}{S_T}, \tag{4.7}$$

where $\bar{SE_T}$ is the average of SE estimates across the replications, and S_T is the population standard error for estimates of the true parameter. However, because the true standard error is unknown, the empirical standard error from replications calculated as the standard deviation of the parameter estimates is used instead. I used the cut-off value of ±0.1 for the acceptable relative bias of SE, based on Hoogland and Boomsma (1998).

4.1.4 Analysis

I generated data for each condition using 1,000 replications. The performance of each approach was evaluated using the criteria listed. I tracked the convergence rates for each approach and performed additional replications so that the simulation results could be calculated based on 1,000 converged replications per condition and model. Monte Carlo standard errors (MCSE) per performance criterion were calculated. Given the finite number of replications, MCSE describes the simulation uncertainty using the SE estimates of the estimated performance (Morris et al., 2019). Specifically, the MCSE of RMSE was calculated using the jack-knife technique (Efron & Stein, 1981).

I employed ANOVA to assess the degree to which coefficient estimates of each method differed from others and determine to what extent each condition affected the performance of each model. I reported the partial eta-squared (η_p^2) effect size associated with the relevant condition being manipulated for each analysis to provide a measure of the practical significance of effects. Based on the ANOVA results, the simulation results were visualized graphically using box plots. The box plots show the median and range of overall performance, including conditions not directly represented in the graph. I used the X-axis as the main condition and the Y-axis as the performance criterion and partitioned the panels by other conditions.

All simulations were performed on `R` 4.2.1 (R Core Team, 2022). I simulated cross-classified data using custom-written `R` code, and estimated CCREM models and the hybrid models using the `lmer()` function of the `lme4` package, which employs constrained optimization of a profiled log-restricted likelihood for REML estimator (Bates et al., 2015). The two-way FE-CRVE model was estimated using the `felm()` function of the `lfe` package (Gaure, 2013a).

4.2 Results

4.2.1 Preliminary Analysis

As a preliminary analysis, I calculated the convergence rates of each method and presented its descriptive statistics in Table 4.3. Most methods have convergence rates close to 1, but the within-cell RE and correlated-cell RE models using the cell-mean centering had considerably lower convergence rates. The ANOVA examining the effect of simulation factors on the convergence rates showed that the method and cell IUCC have a particularly large effect size of 0.828 and 0.693, respectively (see Table 6.1 in Appendix). I presented the convergence rates by methods and cell IUCC in Figure 4.3, which demonstrates that the cell-mean centering methods exhibit the lowest average convergence rate, especially when the cell IUCC is 0. In other words, the low convergence rate observed in the cell-mean centering can be attributed to the challenge of estimating the random effects caused by the low cell IUCC.

Table 4.3

Rate of Convergence of the Coefficient Estimates by Methods

Methods (%)	Min.	Q1	Mean	Median	Q3	Max
Uncentering	93.0	99.5	99.6	100.0	100.0	100.0
Grand-mean centering	93.0	99.5	99.6	100.0	100.0	100.0
Within-RE Model	95.6	99.6	99.8	100.0	100.0	100.0
CRE Model	92.8	99.50	99.6	100.0	100.0	100.0
Within-cell RE Model	31.6	80.7	87.7	94.9	98.7	100.0
Correlated-cell RE Model	22.6	52.6	77.9	89.7	97.05	100.0
FE-CRVE	100.0	100.0	100.0	100.0	100.0	100.0
Hybrid (School FE/ Neighbor RE)	95.2	100.0	100.0	99.9	100.0	100.0
Hybrid (School RE/ Neighbor FE)	92.8	100.0	99.45	100.0	100.0	100.0

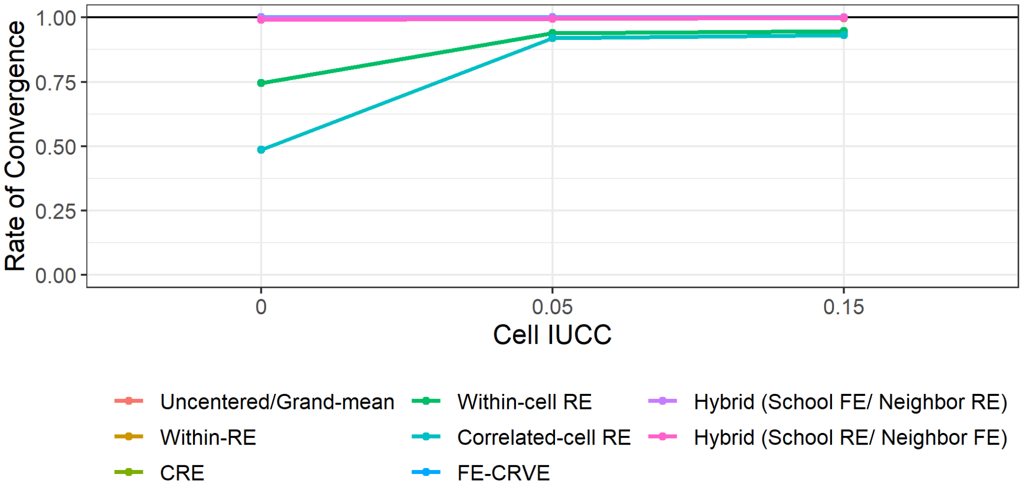

Figure 4.3. Rate of Convergence of the Coefficient Estimates by Methods and Cell IUCC

Further, when there is only one student in a cell, implementing cell-mean centering may fail to converge because subtracting the average of the cell from the raw covariates would result in zero variance. As previously shown in Figure 4.2, the cells with only one student were often observed when the number of clusters condition was large as 70 or 150. However, considering that the effect size of the number of clusters on the convergence rates was relatively small compared to the cell IUCC ($\eta_p^2 = 0.02$, Table 6.1 in Appendix), the low number of students in a cell may have a small impact on the low convergence rate.

Another potential factor contributing to low convergence rates, particularly in the correlated-cell RE model, might be the presence of multicollinearity between the cluster mean and cell mean. Since cells represent a combination of clustering dimensions, there can be a high correlation between the cluster and cell means. I examined the correlation between the cell mean and the school mean as well as the neighborhood mean in the generated data (Table 4.4). Both dimensions exhibited a substantial correlation, with the cell mean demonstrating a slightly stronger correlation with the neighborhood dimension, which suggests a higher chance of multicollinearity. This correlation was influenced by the number of clusters with large effect size: $\eta_p^2 = 0.355$ for the correlation between school mean and cell mean, and $\eta_p^2 = 0.917$ for the correlation between neighborhood mean and cell mean (see Table 6.2 in the Appendix). However, to take this into account, the correlated-cell RE model in the simulation study utilized the adaptively centered cell mean, which is orthogonal to the two cluster means. Thus, the effect of multicollinearity between the cluster means and the cell means on the convergence rate was expected to be minimal. For methods with an insufficient number of replications, additional replications were conducted to reach 1,000.

Table 4.4

Correlation between Cluster Mean and Cell Mean

Clustering Dimension	Min.	1Q	Mean	Median	3Q	Max
School	0.29	0.55	0.57	0.59	0.61	0.86
Neighborhood	0.54	0.62	0.66	0.70	0.79	0.95

I also confirmed whether the number of neighborhoods was generated accurately under the simulation conditions. To this end, I calculated the ratio of the average number of neighborhoods generated to the target number of neighborhoods specified in the conditions (i.e., the average number of neighborhoods is obtained by multiplying the number of schools by 3.5). Based on the results, there were no issues with generating neighborhoods, as the lowest proportion was over 99%. (see Figure 6.1 in the Appendix).

4.2.2 Within-Cluster Effect

When illustrating the simulation results, I first conducted an ANOVA and plotted the results on the box plots for conditions with the largest effect sizes on the X-axis, rows, and columns. In the graph, several pairs of methods that produced the same value of the within-cluster effect were combined: not centering and grand-mean centering, the within-RE model and CRE model using cluster-mean centering, and the within-cell RE model and correlated-cell RE model using cell-mean centering were combined to simplify the graphs.

Parameter Bias

Table 4.5 reports the ANOVA results of relative and absolute parameter bias, indicating that the method had the largest effect size (0.613) on the relative parameter bias, followed by the assumption (0.423) and the coefficient size (0.183). The number of students per school had a medium effect size (0.084). Similarly, the effect size of the method (0.811), assumptions (0.672), and the number of students per school (0.202) were substantial for absolute parameter bias. Thus, Figure 4.4 illustrates the relative and absolute parameter bias as a function of the number of students per school (X-axis), the assumption (columns), and the coefficient size (rows).

The methods that lie between the two dashed lines represent acceptable relative parameter bias (see Figure 4.4A). When all the CCREM assumptions were met, almost all methods exhibited acceptable relative parameter bias. Only cell-mean centering showed a relative parameter bias outside of the acceptable range when the coefficient condition was 0.01, mainly when the number of students per school was as low as 30. The unacceptable relative parameter bias associated with cell-mean centering could potentially be attributed to the MCSE of bias. However, the maximum MCSE was as low as 0.029. Otherwise, the performance of cell-mean centering was relatively decent.

Table 4.5

ANOVA Results on Relative and Absolute Parameter Bias of the Within-Cluster Effects for the Level-1 Covariate

	Relative PB	Absolute PB
Experimental Factors	η_p^2	η_p^2
Method	0.613 (large)	0.811 (large)
CCREM Assumption	0.423 (large)	0.672 (large)
Coefficient	0.183 (large)	0.001
Number of level-2 clusters (schools)	0.002	0.004
Number of level-1 students per school	0.084 (medium)	0.202 (large)
Neighborhood IUCC	0.002	0.008
Cell IUCC	0.004	0.013 (small)

Note. PB indicates parameter bias; Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

When the exogeneity assumption was violated, the performance of not centering, grand-mean centering, and the hybrid model handling schools as FE and neighborhoods as RE was unacceptable. Only cluster-mean centering, cell-mean centering, FE-CRVE, and the hybrid model treating schools as RE and neighborhoods as FE controlled the effects of the exogeneity assumption being violated. Given that the exogeneity assumption was violated in the neighborhood clustering dimension, the latter hybrid model seems to have avoided the correlation between covariates and random effects because of the neighborhood dimension treating the neighborhood dimension with FE.

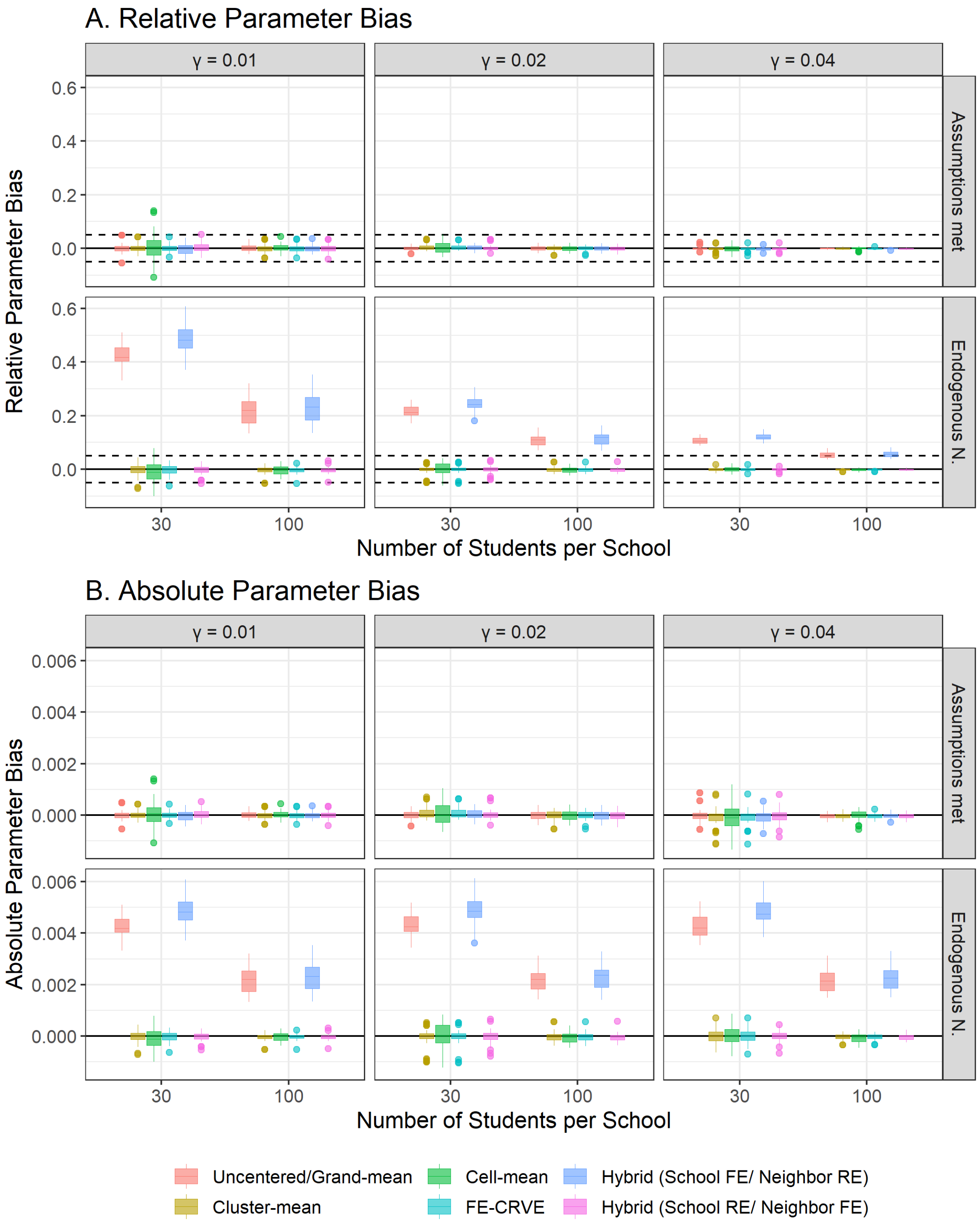

Figure 4.4. Relative and Absolute Parameter Bias of the Within-Cluster Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative and absolute parameter bias was 0.029 and 0.0003, respectively. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

As shown in Figure 4.4B, the absolute parameter bias was close to zero when all the assumptions of CCREM were satisfied. However, when the exogeneity assumption was violated, the bias was pronounced for not centering, grand-mean centering, and for the hybrid model, where schools were modeled as FE and neighborhoods as RE. Specifically, the magnitude of the bias was always slightly larger for the hybrid model than for not centering or grand-mean centering.

The number of students per school substantially affected the bias, with the overall bias approximately ranging from 0.004 to 0.005 when the number of students per school was as small as 30 and about 0.002 when the number of students per school was 100. Also, it is worth noting that the magnitude of the absolute parameter bias remained similar regardless of the coefficient's size, unlike the relative parameter bias. This implies that the bias observed is additive rather than multiplicative.

Root Mean Squared Error

In Table 4.6, the ANOVA results for RMSE shows that the simulation conditions with the largest effect sizes for RMSE were the number of schools (0.938) and the number of students per school (0.913). Although the effect sizes are relatively small, method (0.727), assumption (0.440), and neighborhood IUCC (0.165) also had large effect sizes. Figure 4.5 plots the RMSE using the number of students per school (X-axis), the exogeneity assumption condition (rows), and the number of schools (columns). A lower RMSE is interpreted as better performance.

In Figure 4.5, when the exogeneity assumption was satisfied, use of an uncentered or grand-mean centered covariate in the CCREM had the lowest RMSE, followed by hybrid approaches, cluster-mean centering, and FE-CRVE. The cluster-mean centering and FE-CRVE had identical RMSEs. Cell-mean centering had the highest RMSE compared to other methods, indicating that it loses efficiency compared to the other methods.

When the exogeneity assumption was violated, the unbiased estimators, in-

Table 4.6

ANOVA Results on Root Mean Square Error of the Within-Cluster Effects for the Level-1 Covariate

Experimental Factors	η_p^2
Method	0.727 (large)
CCREM Assumption	0.440 (large)
Coefficient	0.003
Number of level-2 clusters (schools)	0.938 (large)
Number of level-1 students per school	0.913 (large)
Neighborhood IUCC	0.165 (large)
Cell IUCC	0.002

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

cluding cluster-mean centering, FE-CRVE, and a hybrid model that treats schools as RE and neighborhoods as FE, had a lower RMSE than the other methods and appeared to control for endogeneity well. This hybrid model showed slightly better (i.e., lower RMSE) performance among these methods. On the other hand, not centering, grand-mean centering, cell-mean centering, and the hybrid model with the school dimension as FE and neighborhood dimension as RE had relatively higher RMSEs than the first group of methods.

Considering both bias and RMSE results, cell-mean centering specifically had a trade-off that provided a similar degree of inefficiency as the use of the uncentered CCREM but exhibited less bias for the assumption violation.

As expected, all methods generally had lower RMSEs as the number of clusters increased. The number of students per cluster also had an overall positive impact on performance. In other words, the RMSE tended to decrease as the number of students per cluster increased, with the lowest RMSEs observed when the number of students per cluster was around 100.

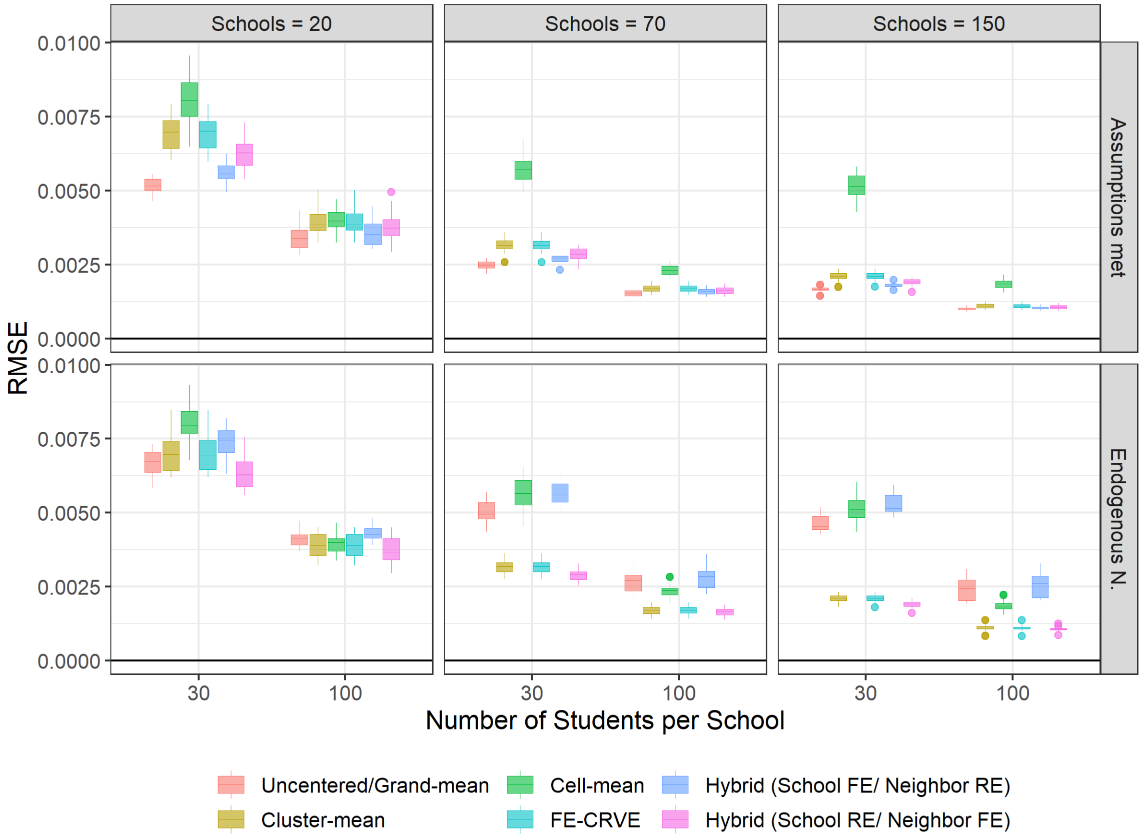

Figure 4.5. Root Mean Square Error of the Within-Cluster Effects for the Level-1 Covariate.
Note. The maximum MCSE was 0.0003. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

Relative Bias of Standard Error

In Table 4.7, I investigated the influence of simulation conditions on the relative SE bias using ANOVA. The results revealed that cell IUCC (0.278) and the number of clusters (0.181) had the most substantial effect sizes, followed by method (0.132) and the number of students per school (0.088). The violation of the assumption did not have any substantial effect on the relative bias of SE (0.006). Thus, in Figure 4.6, the X-axis shows the number of students per school, while rows and columns correspond to the cell IUCC and the number of schools. The dashed lines at 0.9 and 1.1 indicate the acceptable range of relative bias of SE.

Regarding the relative SE bias, only cell-mean centering consistently provided valid inferences. The performance of other methods primarily depended on the cell IUCC. The performance of all methods was acceptable when cell IUCC was 0. However, when the cell IUCC value increased, other models except the cell-mean centering did not capture the additional dependence of the two dimensions and showed unacceptable performance. Even FE-CRVE, which captures the cell-interaction errors, did not perform well when the cell IUCC became large as 0.15.

Table 4.7

ANOVA Results on Relative Bias of Standard Error of the Within-Cluster Effects for the Level-1 Covariate

Experimental Factors	η_p^2
Method	0.132 (medium)
CCREM Assumption	0.006
Coefficient	0.007
Number of level-2 clusters (schools)	0.181 (large)
Number of level-1 students per school	0.088 (medium)
Neighborhood IUCC	0.002
Cell IUCC	0.278 (large)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Further, the impact of cell IUCC on the performance of methods also depended on the number of schools when the cell IUCC is greater than 0.05. For example, when the number of schools was 20, all methods except cell-mean centering estimated SEs within the cut-off value. When the number of schools was large as 150, all methods again performed acceptably, even when cell IUCC remained at 0.05. On the other hand, methods other than cell-mean centering tended to produce underestimated SEs when the number of students per school increased. For example, for larger cell IUCC of 0.15, all methods except cell-mean centering provided underestimated SE when the number of students per school was 100 rather than 30.

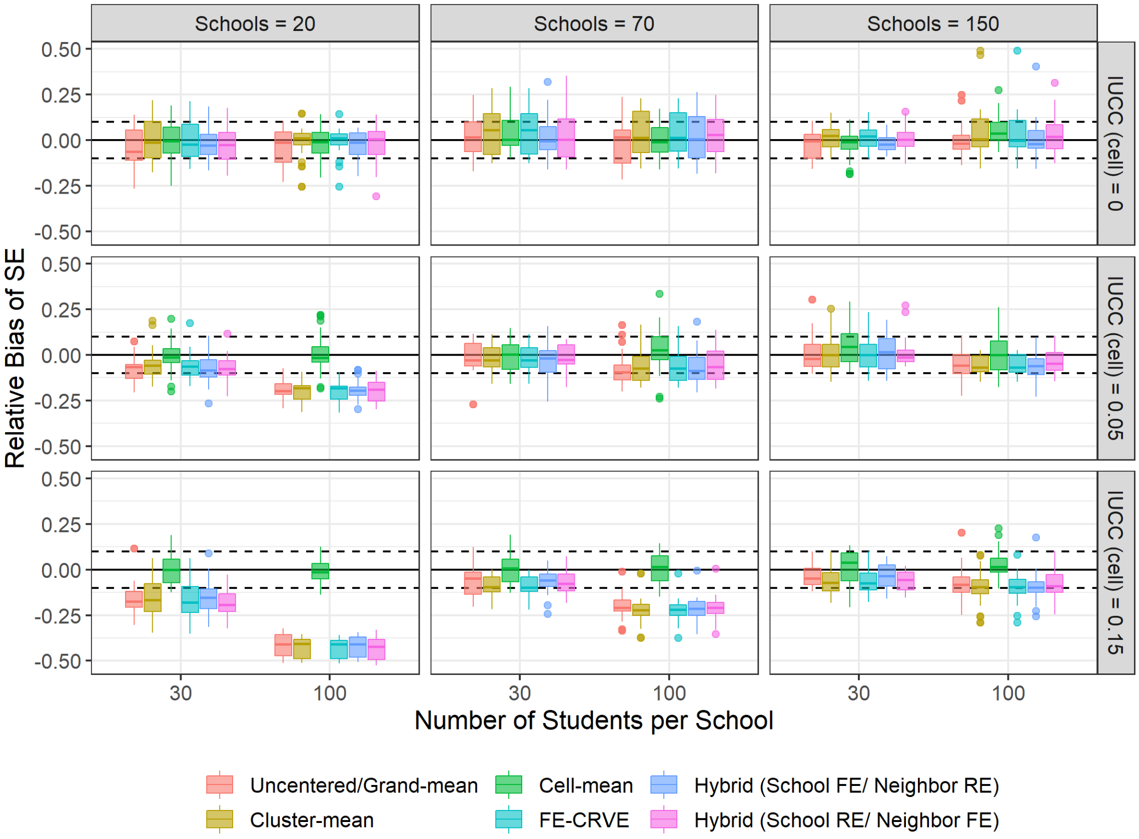

Figure 4.6. Relative Bias of Standard Error of the Within-Cluster Effects for the Level-1 Covariate.

Note. The maximum MCSE was 0.071.

4.2.3 Between-Cluster Effect

Parameter Bias

For between-cluster effects, I considered only the CRE model and the correlated-cell RE model that provide between-cluster effects of level-1 covariates. Table 4.8 displays the ANOVA results of parameter bias for school and neighborhood dimensions. For the school dimension, the methods showed a medium effect on both relative and absolute parameter bias. All simulation factors except cell IUCC exhibited large effect sizes. The number of schools has the greatest impact on the relative (0.891) and absolute parameter bias (0.941). However, the order of effect size was different for the rest of the conditions by relative and absolute parameter bias. For the relative parameter bias, it was followed by the assumption violation (0.611), the number of students per school (0.488), neighborhood IUCC (0.310), and coefficient size (0.303). For the absolute parameter bias, coefficient size (0.903) had the second most substantial effect size, followed by students per school (0.660), assumption (0.651), and neighborhood IUCC (0.504).

Table 4.8

ANOVA Results on Relative and Absolute Parameter Bias of the Between-Cluster Effects for the Level-1 Covariate

	School		Neighborhood	
Experimental Factors (η_p^2)	Relative PB	Absolute PB	Relative PB	Absolute PB
Method	0.071 (medium)	0.138 (medium)	0.000	0.004
CCREM Assumption	0.611 (large)	0.651 (large)	0.959 (large)	0.993 (large)
Coefficient	0.303 (large)	0.903 (large)	0.868 (large)	0.811 (large)
Number of schools	0.891 (large)	0.941 (large)	0.560 (large)	0.902 (large)
Number of students/school	0.488 (large)	0.660 (large)	0.169 (large)	0.579 (large)
Neighborhood IUCC	0.310 (large)	0.504 (large)	0.643 (large)	0.910 (large)
Cell IUCC	0.007	0.008	0.003	0.005

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Regarding the neighborhood dimension, the largest effect was shown for the exogeneity assumption (0.959 for relative bias and 0.933 for absolute bias), possibly due to the dimension's endogeneity. The coefficient size also has a substantial effect size (0.868), followed by the neighborhood IUCC (0.643), the number of schools (0.560), and the number of students per school (0.169) for the relative parameter bias. Further, for the absolute parameter bias, neighborhood IUCC (0.910) had the next largest effect sizes for the neighborhood dimension, followed by the number of schools (0.902), coefficient size (0.811), and the number of students per school (0.579). Considering the ANOVA results, I drew Figures 4.7 and 4.8 with the number of students per cluster on the X-axis and the exogeneity assumption condition, and the number of schools on the rows and columns.

In Figure 4.7, the between-cluster effects for the school dimension exhibited poor relative and absolute parameter bias overall. Only a few cases appeared to fall within an acceptable range, depending on the number of students per school. This parameter bias could be attenuation bias due to using observed mean centering instead of latent mean centering. In other words, the bias in the observed cluster mean due to the insufficient number of students per cluster compared to the cluster population means could have led to the larger attenuation bias.

When the exogeneity assumption was met, the correlated-cell RE model performed a little worse than the CRE model when the number of schools was small. Otherwise, the differences between the CRE and correlated-cell RE models was minimal. The school dimension showed partially acceptable performance only when the number of students per school was large as 100. When the exogeneity assumption was violated, the relative and absolute parameter bias of the school dimensions showed a very slight decrease. This may be because the endogeneity in the neighborhood dimension did not largely affect the between-cluster effect in the school dimension. Further, the relative parameter bias in the school dimension became unacceptable in cases where the number of students per school was 100 and the number of schools was 70.

Figure 4.8 depicts the same plot for the neighborhood dimension. When the exogeneity assumption was met, the neighborhood dimension showed comparable relative and absolute parameter bias to the school dimension's performance. In other words, the parameter bias observed in between-cluster effects in the neighborhood dimensions could also be attributed to attenuation bias, similar to the between-cluster effects in the school dimension. The difference in the CRE and correlated-cell RE models was noticeable only when the number of schools was small. In all other cases, the two models demonstrated similar levels of parameter bias.

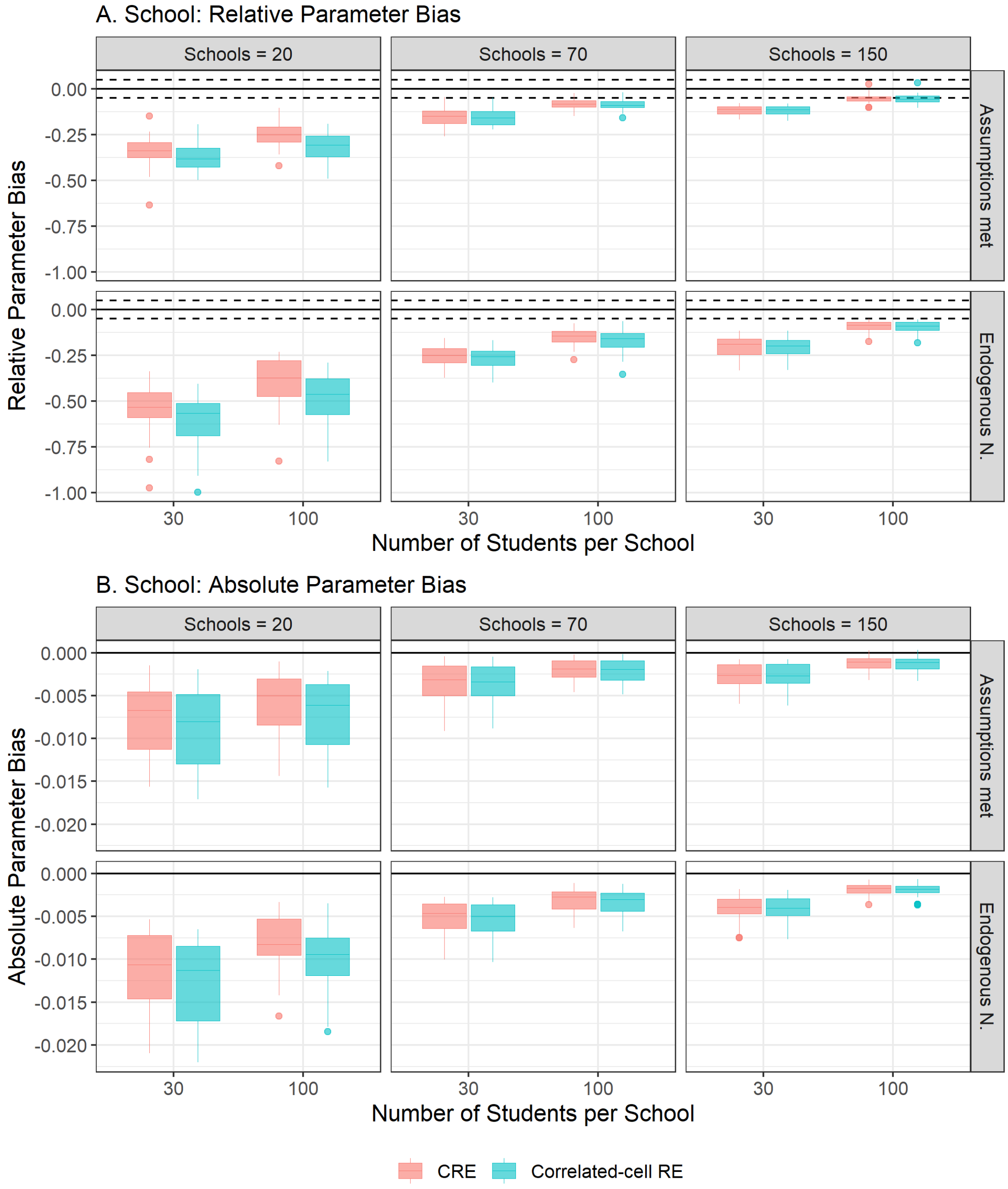

Figure 4.7. Relative and Absolute Parameter Bias of the Between-School Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the school dimension was 0.049, while the maximum MCSE for absolute parameter bias was 0.001. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

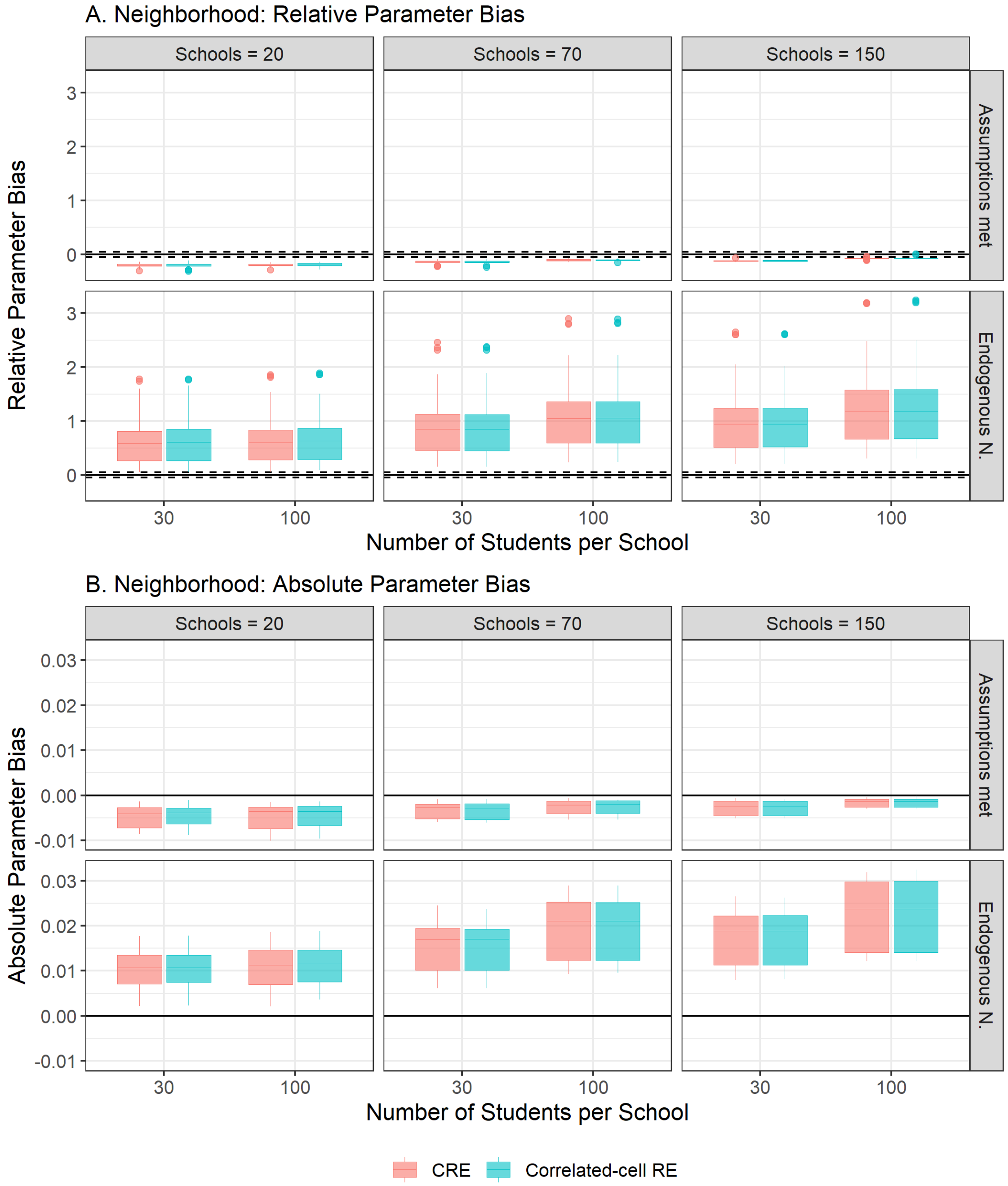

Figure 4.8. Relative and Absolute Parameter Bias of the Between-Neighborhood Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the neighborhood dimension was 0.036, while the maximum MCSE for absolute parameter bias was 0.0004. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

When the endogeneity was introduced into the neighborhood effect, substantial overestimation was observed for the between-cluster effect in the neighborhood dimension, unlike that in the school dimension. This overestimation was unique in the neighborhood dimension, which could be seen as an additive bias in addition to the attenuation bias for the neighborhood coefficients. This tendency was seen across conditions with endogenous neighborhood effects.

Also, as the ANOVA results demonstrated, the number of students per cluster and the number of schools substantially impacted the parameter bias of the between-cluster effect. Specifically, when the number of students per school reached 100 while the number of schools was 150, the relative parameter bias was marginally, but not perfectly, within the acceptable range, except for the neighborhood dimension when the assumption was violated. On the other hand, when the number of students per school was as small as 30, the relative parameter bias exceeded or was below the cut-off criteria.

To explore the results further, I plotted additional simulation conditions with substantial effect sizes. Figures 6.2 and 6.3 (see the Appendix) show the results for the school and neighborhood dimensions, respectively, with neighborhood IUCC plotted in the column instead of the number of schools. Higher IUCCs were associated with more acceptable performance in the school dimension in all conditions. In contrast, higher IUCCs were associated with more overestimation in the neighborhood dimension when the exogeneity assumption was violated.

Moreover, when the exogeneity assumption was not satisfied, the magnitude of the parameter bias in the neighborhood dimension seems substantially affected by the coefficient size (Figures 6.4 and 6.5 in Appendix). Considering the relative parameter bias, the magnitude of this bias decreased as the coefficient size increased. However, the parameter bias of the between-neighborhood effect remained unacceptable.

Root Mean Squared Error

The ANOVA results in Table 4.9 reveal simulation factors that affected the RMSE. All simulation conditions showed a substantial impact on the school dimension. The most considerable effects were observed for the number of schools and coefficient size, with effect sizes of 0.987 and 0.849, respectively. The simulation factors with the following large effect sizes were students per school (0.722), exogeneity assumption (0.445), and neighborhood IUCC (.316), in order. In the neighborhood dimension, the number of schools had a relatively minor impact (0.007). Instead, the conditions with a substantial impact on the neighborhood dimension were assumption violation (0.962), followed by neighborhood IUCC (0. 894), and students per school (0.169).

Table 4.9

ANOVA Results on Root Mean Squared Error of the Between-Cluster Effects for the Level-1 Covariate

	School	Neighborhood
Experimental Factors	η_p^2	η_p^2
Method	0.046 (small)	0.000
CCREM Assumption	0.445 (large)	0.962 (large)
Coefficient	0.849 (large)	0.036 (small)
Number of level-2 clusters (schools)	0.987 (large)	0.007
Number of level-1 students per school	0.722 (large)	0.169 (large)
Neighborhood IUCC	0.316 (large)	0.894 (large)
Cell IUCC	0.150 (large)	0.029 (small)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

In Figure 4.9, I present the RMSE, with the number of students per school on the X-axis and the assumption in the rows. In the columns, the coefficients were drawn for the school dimension depending on the factors that influenced their RMSE (Figure 4.9A), while the neighborhood IUCC condition was placed instead of the coefficient size in the neighborhood dimension (Figure 4.9B).

On the school dimension, the CRE and correlated-cell RE models yielded comparable results in all conditions (see Figure 4.9A). Similar to the results of parameter bias, the violation of the exogeneity assumption did not significantly affect the school dimension. However, the RMSE appeared to decrease as the number of students per school increased. This is likely due to the larger number of students per school, resulting in the model's cluster mean being calculated closer to the population mean. The number of schools also positively impacted RMSE, with increasing school sizes leading to decreased RMSE (see Figure 6.6 in Appendix). However, the effect of the coefficient size was the opposite: as the coefficient size increased, the RMSE also increased, indicating less accuracy.

In the neighborhood dimension, both the CRE and correlated-cell RE models showed similar RMSE (see Figure 4.9B). If all the assumptions were met, the RMSE was slightly lower when the number of students per school was 100 compared to when it was 30, as in the school dimension. However, when the exogeneity assumption was violated, the RMSE became much higher, with great variation across conditions. The impact of the number of students per school reversed, with higher RMSE observed when the number of students per school was as large as 100. Even though the increased number of students per school enables calculating a more accurate cluster mean, it harms the accuracy of the between-cluster effect when the exogeneity assumption is violated. Moreover, as in the school dimension, higher IUCC increased the RMSE.

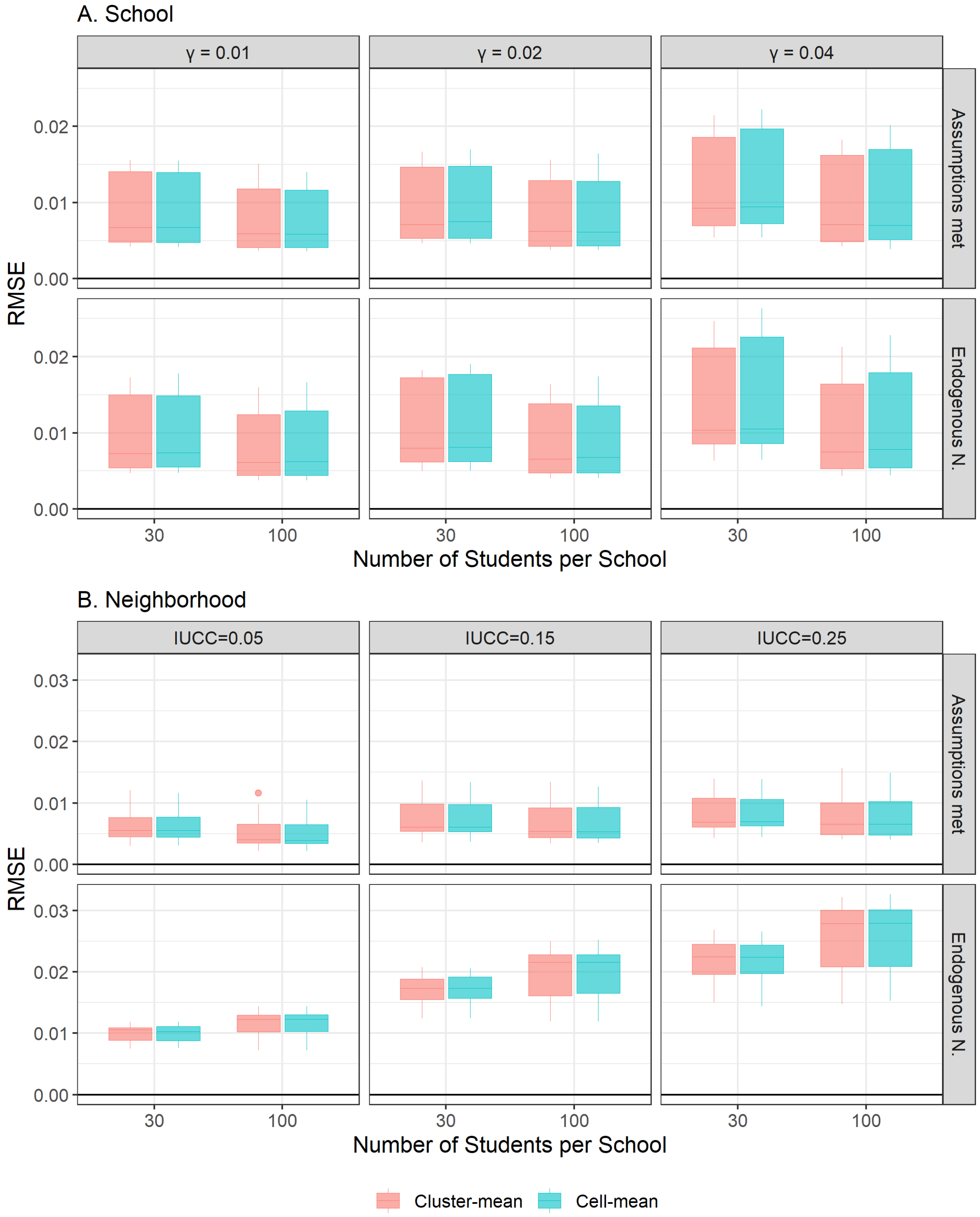

Figure 4.9. Root Mean Squared Error of the Between-Cluster Effects for the Level-1 Covariate.

Note. The maximum MCSE was 0.001 for the school dimension and 0.0004 for the neighborhood dimension. In the neighborhood dimension (B), the IUCC condition in columns indicates neighborhood IUCC. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

Relative Bias of Standard Error

Table 4.10 presents the ANOVA results for the relative bias of SE. Unlike the parameter bias and RMSE, the relative bias of SE did not appear to be substantially affected by the simulation conditions. For the school dimension, the effect sizes of the number of schools (0.027), the coefficient size (0.036), and the number of students per school (0.015) were very small. The exogeneity assumption had a small effect on the neighborhood dimension (0.049), along with the coefficient size (0.072), number of schools (0.050), and number of students per school (0.013), in effect size order.

Table 4.10

ANOVA Results on Relative Bias of Standard Error of the Between-Cluster Effects for the Level-1 Covariate

	School	Neighborhood
Experimental Factors	η_p^2	η_p^2
Method	0.003	0.006
CCREM Assumption	0.002	0.049 (small)
Coefficient	0.036 (small)	0.072 (small)
Number of level-2 clusters (schools)	0.027 (small)	0.050 (small)
Number of level-1 students per school	0.015 (small)	0.013 (small)
Neighborhood IUCC	0.005	0.005
Cell IUCC	0.000	0.004

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

In Figure 4.10, I plotted the number of schools on the X-axis. The exogeneity assumption and coefficient sizes were plotted in the rows and the columns, respectively. Overall, the relative bias of SEs was within an acceptable range, except for a few instances. In the school dimension in Figure 4.10A, the CRE model performed well when the exogeneity assumption was met, except for cases where the number of schools was as small as 20, and the coefficient size was larger than 0.02. The performance of the correlated-cell RE model was relatively decent regardless of the number of schools. The violation of the exogeneity assumption had little impact on the performance of the school dimension's between-cluster effect. Although some of the CRE model results showed an unacceptably underestimated SE when the number of schools was as small as 20, the rest demonstrated an acceptable bias of SE.

In the neighborhood dimension (see Figure 4.10B), the relative bias of SE was acceptable depending on the coefficient size. When the coefficient sizes were 0.01 and 0.02, violation of the assumption had little effect on the results, and most of the results fell within an acceptable range. However, when the coefficient size was 0.04 and the exogeneity assumption was met, the relative bias of SE in the correlated-cell RE model was overestimated and not acceptable. Conversely, when the coefficient size was 0.04 and the exogeneity assumption was violated, more cases showed an underestimated SE, especially when the cluster size was less than 70. The number of schools also influenced the relative bias of SE. When the exogeneity assumption was met, a greater number of schools resulted in an unacceptable bias of SE. On the other hand, a smaller number of schools showed the biased SE below the cut-off value.

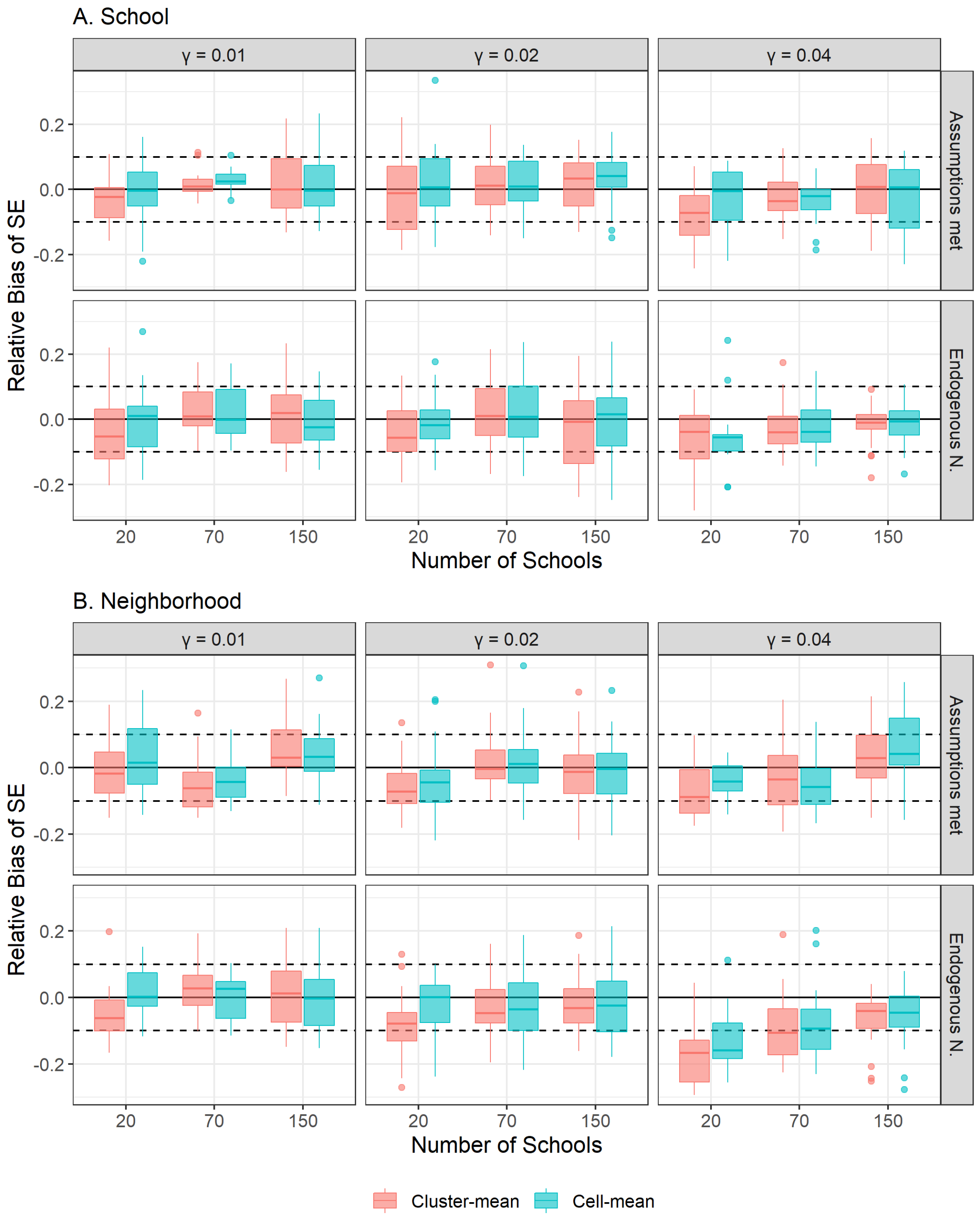

Figure 4.10. Relative Bias of Standard Error of the Between-Cluster Effects for the Level-1 Covariate.

Note. The maximum MCSE was 0.066 for the school dimension and 0.063 for the neighborhood dimension. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

Chapter 5

Discussion

The CCREM has been used to analyze cross-classified data. However, research has not assessed the robustness of results when the exogeneity assumption is violated. In this study, I focused on the exogeneity assumption. Using a systematic review, I examined how this assumption has been addressed in previous studies using the CCREM. Specifically, I examined ten years of CCREM studies outlining the model's characteristics, including the testing of the exogeneity assumption and the use of covariate centering. This study is the first systematic review of the methodology used in applied CCREM research based on analogous work in previous systematic reviews of HLM research (Antonakis et al., 2021; Dedrick et al., 2009; Luo et al., 2021).

I also conducted a simulation study to provide researchers with alternatives that account for the possibility of violating the exogeneity assumption. I first presented the adaptive centering method proposed in Raudenbush (2009) and extended it to the CRE model, enabling the estimation of both within- and between-cluster effects of the covariate. Then, considering that interactions between dimensions can occur for cross-classified data (Shi et al., 2010), I proposed a cell-mean centering approach, including the within-cell RE and correlated-cell RE model.

FE approaches can provide alternative methods for handling endogeneity in cross-classified data. Cameron et al. (2011) proposed two-way FE-CRVE providing robust estimators when focusing on the within-cluster effects of level-1 covariates. Lee and Pustejovsky (2023) have shown that FE-CRVE outperforms CCREM using the raw value of the covariates when the exogeneity assumption is violated. In this paper, I adopted CCREM with adaptive centering, an equivalent estimator to the FE-CRVE. Finally, I proposed a hybrid method that models the cross-classified data using FE in one dimension and RE in one dimension.

My simulation study centered on a comparative evaluation of the above methods under conditions with imbalanced cross-sectional data. I focused on the within- and between-cluster effect coefficients of level 1 covariates. This section will summarize the study's findings and discuss the limitations and potential extensions for future research. Additionally, I will discuss the implications of data analyses by applied researchers.

5.1 Summary of Findings

5.1.1 Systematic Review

In the systematic review, I examined how exogeneity assumptions have been addressed in previously published CCREM studies and whether the corresponding alternatives (e.g., cluster-mean centering) have been selected accordingly. Considering study characteristics, CCREM was more frequently used in cross-sectional studies where individuals are included in multiple clusters rather than in longitudinal studies. The size of the cross-classified datasets varied widely, ranging from fewer than 20 clusters in the smallest study to nearly 1400 clusters in the largest study.

The multilevel structure of the cross-classified data was typically reflected in the model, although some simplifications were often made. For example, when the cross-classified data had more than three levels, researchers sometimes omitted one of the higher levels at their discretion, resulting in a model with fewer levels. A systematic review of HLM has also reported similar scenarios (Luo et al., 2021). Simple two-way models were most often used when there were more than two clustering dimensions in HLM data (e.g., age, period, and cohort).

The assumptions underlying CCREM have not been adequately assessed in practice. Particularly, the exogeneity assumption appeared to be neglected in the CCREM literature. Very few studies mentioned this assumption or considered the consequences if it was violated. Furthermore, fewer than 2% of CCREM studies attempted to address or mitigate the impact of endogeneity using sensitivity analyses or centering techniques. This tendency was comparable to results found in the systematic review of HLM, where only 4% of studies evaluated the exogeneity assumption (Antonakis et al., 2021). Compared to the normality and homoscedasticity assumptions, the low percentage of studies considering the exogeneity assumption may be due to a lack of clear methods for evaluating this assumption with the CCREM. Also, as with other kinds of models, it can be challenging for researchers to identify the potentially omitted variables. These findings imply that a substantial number of studies may have overlooked the potential consequences of violating the exogeneity assumption.

In terms of centering, 36% of the CCREM studies were found to use grand-mean centering for level-1 or higher level covariates. However, grand-mean centering alone is insufficient for remedying violations of the exogeneity assumption. The percentage of studies implementing cluster-mean centering, which reduces the impact of this assumption violation, was 10%, less common than grand-mean centering. Compared to HLM studies that used cluster-mean centering in 28% of selected studies (Luo et al., 2021), a smaller percentage of CCREM studies have utilized cluster-mean centering.

It was surprising that none of the CCREM studies that used the within-RE model and cluster-mean centering correctly centered their data on multiple cross-classified dimensions. Calculating cluster-mean-centered covariates has often considered only one clustering dimension. Such cluster-mean centering can pose problems in CCREM as it fails to account for potential endogeneity in the other clustering dimension. If the remaining clustering dimension has endogeneity, it can lead to a violation of the exogeneity assumption and undermine the robustness of the results.

Regarding the software used in the literature, CCREM studies used more diverse software than HLM literature (Luo et al., 2021). `HLM` software was found to be used in about half of applied HLM studies, whereas `MLwiN` and `R` were most commonly used in about 40 percent of the applied CCREM studies that were reviewed. Further, the most frequent estimation method use in CCREM studies was the Bayesian approach, which provides more reasonable results with the high complexity of the variance structure of the model (Baldwin & Fellingham, 2012), followed by ML and REML.

In conclusion, the systematic review revealed that the assessment and handling of assumptions in CCREM studies, particularly the exogeneity assumption, was lacking. Only a very small proportion of studies considered the consequences of violating this assumption and employed techniques to address it. Although centering techniques like cluster-mean centering were used in a few studies, the practice of centering covariates on multiple cross-classified dimensions was surprisingly absent. This practice of cluster-mean centering could be due to the lack of clear guidelines within the context of CCREM. Also, this oversight can lead to inadequate handling of potential endogeneity in cross-classified data, compromising the robustness of the results. Thus, this study emphasized bridging this gap by proposing alternative approaches that provide more accurate statistical inferences when confronted with the potential violation of the exogeneity assumptions.

5.1.2 Simulation Study

Considering the issues in the systematic review, the simulation study evaluated various proposed techniques for managing violations of the exogeneity assumption in cross-classified data. I generated unbalanced cross-classified data and compared the performance of different methods, including not centering, grand-mean centering, cluster-mean centering, cell-mean centering, FE-CRVE, and two hybrid models incorporating FE and RE on each clustering dimension, respectively.

Within-Cluster Effect

The simulation study results demonstrated that the best method for estimating the within-cluster effects for the level-1 covariates depended on whether the exogeneity assumption was met. When the exogeneity assumption was met, all the methods showed acceptable levels of relative parameter bias and negligible absolute parameter bias with a decent relative SE bias. In RMSE, however, there was a disparity observed across the methods. Uncentered CCREM and grand-mean centering consistently demonstrated the best performance, followed by hybrid models, cluster-mean centering, and FE-CRVE with small differences. As expected, the performance of cluster-mean centering and FE-CRVE was the same. On the other hand, cell-mean centering showed the highest RMSE, indicating the lowest accuracy even when the exogeneity assumption was met. Specifically, the performance of cell-mean centering was worse when the number of schools and students per school was lower.

When the exogeneity assumption was violated, there was a noticeable decline in the performance of not centering, grand-mean centering, and the hybrid model treating the school clustering dimension as FE and the neighborhood dimension as RE, in addition to the cell-mean centering. In particular, CCREM with uncentered and grand-mean-centered covariates resulted in unacceptably biased coefficient estimates because these approaches provided the pooled effect of the covariates. The pooled effect is likely to be biased due to the bias in the between-cluster effect when the exogeneity assumption was not met.

The hybrid model, which specified the school dimension as FE and the neighborhood dimension as RE, also produced unacceptable parameter estimates. By specifying neighborhood dimensions as RE without controlling for its potential endogeneity, the between-cluster effects in the neighborhood dimension may have been confounded with the within-cluster effects. As a result, the estimated coefficient for within-cluster effects might have been overestimated compared to other methods.

Even though the exogeneity assumption was violated, cluster-mean centering, FE-CRVE, and the hybrid model, where the school dimension was treated as RE and the neighborhood dimension was treated as FE, consistently performed well as measured using the RMSE, particularly when the number of clusters exceeded 70. Among these methods, the hybrid model provided a slightly more accurate estimator than FE-CRVE and cluster-mean centering, with the latter two methods exhibiting the same RMSE. These findings suggest that when the hybrid model addresses the endogeneity in the correct clustering dimension through FE, the accuracy of the methods improves compared to the FE-CRVE and cluster-mean centering, even when the remaining clustering dimension was modeled with RE.

Cell-mean centering consistently exhibited poor performance in terms of RMSE when the exogeneity assumption was violated. The low accuracy in cell-mean centering can be attributed to its higher variance in the estimator, although the estimator was unbiased. Not centering, grand-mean centering, and the hybrid model that used FE on the school dimension and RE on the neighborhood dimension additionally showed higher RMSE when the endogeneity was introduced in the neighborhood dimension. This implies that uncontrolled endogeneity in neighborhood dimension diminishes the accuracy of the estimator in these methods.

Lastly, the performance of the relative SE bias was not strongly related to the violation of the exogeneity assumption. Instead, the relative SE bias varied as a function of the cell IUCC, the number of schools, and the number of students per school. When the cell IUCC was zero, indicating no cell-interaction random effect, all models demonstrated a favorable relative SE bias. However, when the cell IUCC increased to 0.15, only cell-mean centering consistently estimated an acceptable SE under all conditions. When the number of schools was small and the number of students per school was large, all models except for cell-mean centering showed unacceptably low SEs. It is worth noting that despite accounting for the cell-interaction random effect, FE-CRVE also exhibited underestimated SEs, similar to other models. In other words, the performance of FE-CRVE and cell-mean centering may differ in estimating

SE, even when they account for cell-interaction random effects.

Between-Cluster Effect

The between-cluster effect was available only in the CRE and correlated-cell RE models. Although the CRE model consistently provided slightly better relative and absolute parameter bias than the correlated-cell RE model, the difference between the two centering methods was minimal when the exogeneity assumption was met. However, when the exogeneity assumption was violated, the parameter estimates of the CRE and correlated-cell RE model were found to be slightly more downward biased for school clustering dimensions. In contrast, in the neighborhood clustering dimension, the CRE and correlated-cell RE models exhibited substantially overestimated parameter estimates when endogeneity was present.

The number of students per school greatly affected the relative and absolute parameter bias for between-cluster effects. Specifically, in the school dimension, the between-cluster effects in the two methods performed better as the number of students per school increased. This impact was also seen in the neighborhood dimension when the exogeneity assumption was met. Considering that the greater number of students per school can result in cluster sample mean values closer to the population mean, this effect seems reasonable. In other words, there could be an attenuation bias in the observed parameter bias based on the number of students sampled per school. Thus, in the data condition with a small number of students per school, an alternative approach, such as latent-mean centering, has been suggested to reduce the attenuation bias in estimating between-cluster effects.

However, unlike the scenario in which increasing the number of students per school had mitigated bias from other conditions, the increased number of students per school did not improve the parameter bias when endogeneity existed in the neighborhood dimension. This indicates that the bias due to endogeneity may have been larger than the attenuation bias when the endogeneity was presented in the neighborhood dimension.

Regarding RMSE, both methods demonstrated similar performance, although the CRE model often showed slightly better RMSE. The impact of the number of students per school on the RMSE was similar to that on the parameter bias. In the school dimensions, or when the exogeneity assumption was met in the neighborhood dimension, a larger number of students per school resulted in a more accurate parameter estimate. However, when the exogeneity assumption was violated, a greater number of students per school led to higher RMSE, indicating lower accuracy. The neighborhood IUCC was another critical factor negatively associated with the RMSE in the neighborhood dimension. In cases where endogeneity was present, lower IUCC provided a better RMSE for the parameter estimate.

For the relative bias of SE, both the CRE and correlated-cell RE models generally showed acceptable performance. However, when the exogeneity assumption was violated, the relative bias of SE was found to be unacceptable in the neighborhood dimension when the number of schools was less than 70. This tendency was particularly noticeable when the coefficient size was large as 0.04.

5.2 Limitations and Future Directions

Several limitations of this simulation study should be noted for future research. First, I primarily focused on cases where the exogeneity assumption is violated for a specific clustering dimension. However, as briefly mentioned in the Data Generation section, there is no guarantee that the exogeneity assumption is violated in only one clustering dimension. Instead, it can be violated in multiple clustering dimensions or even in interaction terms between the clustering dimensions. In particular, the violation of the exogeneity assumption in interaction terms is a unique scenario that can only occur in cross-classified data. In such cases, cell-mean centering can offer a more robust performance than cluster-mean centering by simultaneously accommodating endogeneity in multiple clustering dimensions and interaction terms. On the other hand, the cluster-mean centering recommended in this study is limited to addressing endogeneity only within the clustering dimension that includes the cluster-mean-centered covariate. Thus, to assess the impact of endogeneity when the exogeneity assumption is not limited to a single clustering dimension, further simulation studies on an extended range of the exogeneity assumption are needed.

Further, this study did not address the other assumptions in CCREM, namely the homogeneity assumption, the linearity assumption, and the normality assumption. In the case of the normality assumption, the HLM, especially the random intercept model, is robust to the non-normal random effects (McCulloch & Neuhaus, 2011). For the homogeneity assumption, Hedeker et al. (2008) and Lee and Nelder (2006) proposed a mixed-effects location-scale model that relaxes the homogeneity assumption by modeling variance differences (for applications, see Leckie et al., 2014; Rast et al., 2012). In a mixed-effects location-scale model, the homoscedasticity assumption posed at level-1 was relaxed by modeling it as a log-linear function of the level-1 and level-2 covariates and associated level-2 random effects. Brunton-Smith et al. (2012) later extended this model and showed how it could be used for cross-classified data.

To the best of my knowledge, however, the performance of mixed-effects location scale models on cross-classified data has not been compared to other models suggested in this study. For example, the alternatives of using OLS or FE using CRVE also avoid the homogeneity assumption when analyzing cross-classified data. Future research could compare two-way mixed-effects location-scale models to other CRVE methods to determine their differences and trade-offs. Also, a future study can further examine whether CCREM can simultaneously avoid the exogeneity and homogeneity assumptions by incorporating covariate centering into these models.

Second, the CRE and correlated-cell RE models employed an observed mean centering approach to calculate the cluster means. As discussed earlier, the simulation results demonstrated that increasing the number of students per school reduced bias when estimating between-cluster effects, suggesting that this bias may be attenuation bias. Unless all students in the school are sampled, the observed mean introduces measurement error in the calculating sample cluster means, even with a large number of students per school. In such cases where the number of sampled units from the true population is insufficient (i.e., sampling ratio), the latent mean centering approach was recommended as it accounted for measurement error while estimating the contextual effects (Lüdtke et al., 2008). The effectiveness of latent mean centering in hierarchical data also depended on the aggregation process (Grilli & Rampichini, 2011; Lüdtke et al., 2008).

The sampling ratio may vary across clustering dimensions in cross-classified data, potentially resulting in an increased discrepancy between the observed mean and latent mean centering. In future studies exploring latent mean centering, incorporating the sampling ratio as a simulation condition might be essential. This can be achieved by generating a finite number of units in each clustering dimension and then manipulating the sampling ratio to determine the number of students per school. In this case, it would be critical to manipulate the sampling ratio on multiple cluster dimensions at the same time to consider a cross-classified data structure. Moreover, adaptive centering using the FWL theorem used in this study also has never been compared to latent mean centering. The adaptive centering technique accommodates imbalanced data and can provide a consistent estimate that takes into account the small number of samples present in each cell. Thus, future research may compare and evaluate the performance of latent mean centering and adaptive centering considering different sampling ratios of clusters.

Third, the simulation study used a large enough number of cluster conditions to obtain acceptable performance, similar to the approach in large-scale survey data. In other words, this simulation study did not have to consider small sample corrections. However, it is typical for researchers to have financial constraints and practical limitations in collecting a sufficient number of clusters (Maas & Hox, 2005). For example, in the results of the systematic review, the Q1 of the number of clusters recorded with the smaller of the two clustering dimensions was 21 (see Table 3.2).

However, as shown in the simulation results, the number of clusters substantially impacted the performance, including the parameter bias of the between-cluster effect, RMSE, and the relative SE bias for both within- and between-cluster effect estimates. In this sense, exploring the robustness of CCREM studies with fewer clusters would be necessary. For example, future studies can address small-sample corrections, such as the Kenward-Roger correction, proposed for linear mixed models (Kenward & Roger, 1997). Investigating the impact of small-sample corrections on the performance of CCREM estimates could provide practical solutions for applied researchers.

Fourth, the random intercept model, the simplest form of CCREM, was used to generate data and estimate the CCREM in the simulation study. The systematic review results indicated that more than 80 percent of recent CCREM studies used the random intercept model. Thus, the random intercept model might be a good starting point for comparing alternatives when the exogeneity assumption is violated. In practice, however, the CCREM can be more complex. For instance, modeling a random slope for a covariate allows researchers to test the assumption that the slope varies between clusters.

Further, the performance of a random intercept model and a random slope model of CCREM may not be comparable. Lee and Pustejovsky (2023) revealed that performance becomes worse when random slopes are misspecified in CCREM than when random slopes are correctly modeled, even when all other assumptions are met. In light of this, even when cluster-mean centering is employed, if the random slope is not correctly modeled, it may not perform as well as other alternatives, such as FE-CRVE. Thus, future research should investigate the strengths and limitations of available centering methods when more diverse random effect structures are utilized.

Another aspect that this simulation study did not consider was the estimation of the variance component. Instead, this study mainly focused on covariates' coefficient estimates (i.e., fixed effects) to assess the performance of different centering methods. However, in the context of RE models, including CCREM, the variance component is an essential indicator to be reported. Specifically, the variance component is used to calculate the IUCC (ICC in HLM), which provides information on the degree of clustering in each dimension. If the variance component is biased (e.g., overestimated) in specific clustering dimensions due to the endogeneity, for example, the results of IUCC might not be accurate. In particular, considering the variance components in the empirical examples I have described above, the value of the variance component varied depending on different centering approaches, the inclusion of cluster means, or random interaction effects in the RE models. Therefore, evaluating the unbiased estimation of the variance component in each model and different conditions might be as critical as assessing the bias of the coefficient estimates.

Fifth, when specifying the hybrid model in the simulation study, I used an uncentered covariate instead of employing centering methods. The uncentered covariate was used to examine the performance degradation of the coefficient of the within- and between-cluster effect when the endogeneity was not properly handled in one clustering dimension. However, the cluster-mean-centered covariate could be used instead of the raw covariate in the hybrid model to handle the potential endogeneity. Using a cluster-mean-centered covariate in the hybrid model is expected to eliminate potential correlations between the covariate and the random effects that cannot be accounted for with fixed effects for one dimension alone. Although this approach would have added complexity to the model, using cluster-mean centering would enable the hybrid model to address the endogeneity issue in the dimension specified with RE while providing coefficients comparable to those obtained in an FE model. Future research should include these extensions of the hybrid model to evaluate the performance compared to different alternatives.

Finally, the number of students per cell should be considered as an additional condition in future simulation studies. While the number of students per school is a typical indicator of the size of the level-1 unit in hierarchical data, cross-classified data requires considering both the number of students per cluster and the number of students per cell as relevant conditions. In this study, I addressed the potential effect of the number of students per cell in Appendix 6.3. The ANOVA results demonstrated that the number of students per cell exhibits some impact on the performance of the within and between-cluster effects. However, the simulation conditions were not fully exhaustive because the number of students per cell was an ad-hoc calculation based on the existing other conditions (i.e., the number of clusters and the number of students per school). A more systematic approach is required to obtain a comprehensive understanding of how the number of students per cell influences the performance of within and between-cluster effects. For example, in cases where the number of schools is 20 and the number of students per school is 30, it would be essential to incorporate conditions representing small and large numbers of students per cell, respectively, to obtain accurate insights into the impact of this variable.

5.3 Implications for Data-Analysis Practice

Based on the findings of this study, several suggestions for future research can be made. First, it is essential to improve researchers' awareness of the exogeneity assumption, the impact of ignoring it and how best to handle it. The exogeneity assumption has a substantial impact on the performance of the within- and between-cluster coefficient estimation in CCREM and should be considered and handled with the same level of priority as other assumptions, such as normality, linearity, and homoscedasticity.

To address this issue of exogeneity assumptions not being considered, I suggest a sensitivity analysis involving multiple models discussed in the simulation study to the same cross-classified data to assess how the coefficients' values might vary. The sensitivity analysis requires two models: one that provides consistent within-cluster effects (e.g., cluster-mean centering or cell-mean centering, or FE-CRVE) and a pair of hybrid models that alternately assumes RE and FE for each cross-classified dimension. The within-cluster effects reported by the aforementioned models will serve as a benchmark for evaluating the endogeneity assumption. If RE is applied to the dimension where the exogeneity assumption holds, the hybrid model will provide comparable values to the within-cluster effects obtained using the model with within-cluster effects. However, if RE is applied to the dimension where endogeneity is present, the hybrid model will likely provide different within-cluster effect estimate values. By comparing these coefficients from hybrid models to one of the models above, researchers can determine whether the exogeneity assumption is met and which dimensions may violate the exogeneity assumption.

The key distinction of this sensitivity analysis from the Hausman test is that it can reveal the specific clustering dimension exhibiting endogeneity. Although the Hausman test provides a general assessment of endogeneity, it does not identify the specific clustering dimension(s) that involves endogeneity. Further, by comparing the differences in the baseline within-cluster effects across dimensions, researchers can determine and compare the extent of endogeneity in each dimension. This enables researchers to better understand the source of endogeneity and effectively choose methods to address the endogeneity within specific clustering dimensions.

Upon confirming the presence of endogeneity in the analysis, I suggest that researchers use the methods proposed in this study. I have included the code for each method in the Appendix to facilitate the implementation of the proposed methods by applied researchers. Specifically, employing cluster-mean centering and FE-CRVE is recommended. In cases where the dimension with endogeneity is identified, I recommend using a hybrid model specifying FE on that dimension. Simulation results demonstrated that the cluster-mean centering and FE-CRVE clearly offer more robust and reliable within-cluster effects. The hybrid model provided an unbiased estimator only when the FE was applied to the clustering dimension where the endogeneity existed. These models generate unbiased within-cluster effects and exhibit decent

RMSEs irrespective of whether the exogeneity assumption is violated. For cases where the number of students per cluster is large or the cell IUCC is higher than 0.05, I advise researchers to interpret significance test results with caution, as the SE is underestimated in all models except those employing cell-mean centering.

The performance of the between-cluster effects is comparatively poor in both the CRE model and the correlated-cell RE model, in contrast to the within-cluster effects. The between-cluster effects can introduce a substantial bias, regardless of whether the exogeneity assumption is violated. Unbiased between-cluster effects can only be partially obtained in ideal scenarios where the number of schools and the number of students per school are large when the exogeneity assumption is not violated. Thus, I suggest considering between-cluster effects only under such a data constellation. If the data fail to meet these conditions (a large number of clusters and level-1 units per cluster), it would be preferable to report an estimate of the between-cluster effect while acknowledging its potential bias.

However, it is worth noting that these results are unique to the cross-classified data generated under the assumption of the presence of cell-interaction effects and the cases where the applied centering methods are based on a random-intercept CCREM. Also, the endogeneity was generated only in one specific clustering dimension (here, neighborhood). In other words, the performance of the methods discussed above cannot be perfectly generalized to all CCREM scenarios. Particularly when dealing with situations where cell-interaction effects are truly negligible or when estimating more complex CCREMs that incorporate random slopes, the performance of the methods may change. When the endogeneity is placed in multiple clustering dimensions or even the interaction terms of the clustering dimensions, the performance of the methods might differ. For example, cell-mean centering might perform better than presented in this simulation study. Moreover, the contextual effect was not directly evaluated in this simulation study. However, considering contextual effects represent the difference between within- and between-cluster effects, the performance in estimating contextual effects is expected to align with that of between-cluster effect estimation. I hope that researchers recognize the importance of the exogeneity assumptions in analyzing cross-classified data and consider the suggestions above to widen the range of alternatives they can employ.

Chapter 6

Appendix

6.1 Cell-Mean-Centered Covariate Considering Cell-Interaction Effect

Considering balanced cross-classified data as a two-way factorial design, a population cell-interaction effect, ϕ_{jk}, is an effect that cannot be accounted for by the overall effect, μ, and main effects for factors, $(\mu_j - \mu)$ and $(\mu_k - \mu)$:

$$\begin{aligned} \phi_{jk} &= (\mu_{jk} - \mu) - (\mu_j - \mu) - (\mu_k - \mu) \\ &= \mu_{jk} - \mu_j - \mu_k + \mu, \end{aligned} \tag{6.1}$$

where μ_{jk} is the cell population mean for the combination of students in school j and neighborhood k (Stevens, 2007, Ch. 4). The estimated cell-interaction effect $\hat{\phi}_{jk}$ is expressed as

$$\begin{aligned} \hat{\phi}_{jk} &= (\bar{X}_{jk} - \bar{X}) - (\bar{X}_j - \bar{X}) - (\bar{X}_k - \bar{X}) \\ &= \bar{X}_{jk} - \bar{X}_j - \bar{X}_k + \bar{X}, \end{aligned} \tag{6.2}$$

where $\bar{X}_{jk}$ indicates the cell sample mean. Rearranging Equation 6.2, the cell mean in the balanced design implies

$$\bar{X}_{jk} = \bar{X}_j + \bar{X}_k + \hat{\phi}_{jk} - \bar{X}. \tag{6.3}$$

Thus, the cell-mean-centered covariate is calculated as

$$\begin{aligned} X_{cell} &= X_{i(jk)} - \bar{X}_{jk} \\ &= X_{i(jk)} - \bar{X}_j - \bar{X}_k - \hat{\phi}_{jk} + \bar{X}. \end{aligned} \tag{6.4}$$

In balanced cross-classified data, considering that the cluster-mean-centered covariate was calculated as $X_{cluster} = X_{i(jk)} - \bar{X}_j - \bar{X}_k + \bar{X}$ (Raudenbush, 2009), Equation 6.4 further incorporates the cell-interaction effects, $\hat{\phi}_{jk}$, of a covariate within the model. In other words, using cell-mean centering has implications for considering cell-interaction effects, $\hat{\phi}_{j\times k}$, when calculating specific covariate effects. However, this interpretation is only valid for data in a balanced design.

6.2 Simulation Study Results

Table 6.1

ANOVA Results on Rate of Convergence by Methods

Experimental Factors	η_p^2
Method	0.828 (large)
CCREM Assumption	0.000
Coefficient	0.013 (small)
Number of level-2 clusters (schools)	0.020 (small)
Number of level-1 students per school	0.000
Neighborhood IUCC	0.001
Cell IUCC	0.693 (large)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Table 6.2

ANOVA Results on Correlation between Cluster Mean and Cell Mean

	School	Neighborhood
Experimental Factors	η_p^2	η_p^2
Method	0.000	0.000
CCREM Assumption	0.000	0.000
Coefficient	0.355 (large)	0.917 (large)
Number of level-2 clusters (schools)	0.005	0.081 (medium)
Number of level-1 students per school	0.000	0.000
Neighborhood IUCC	0.000	0.000
Cell IUCC	0.000	0.000

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

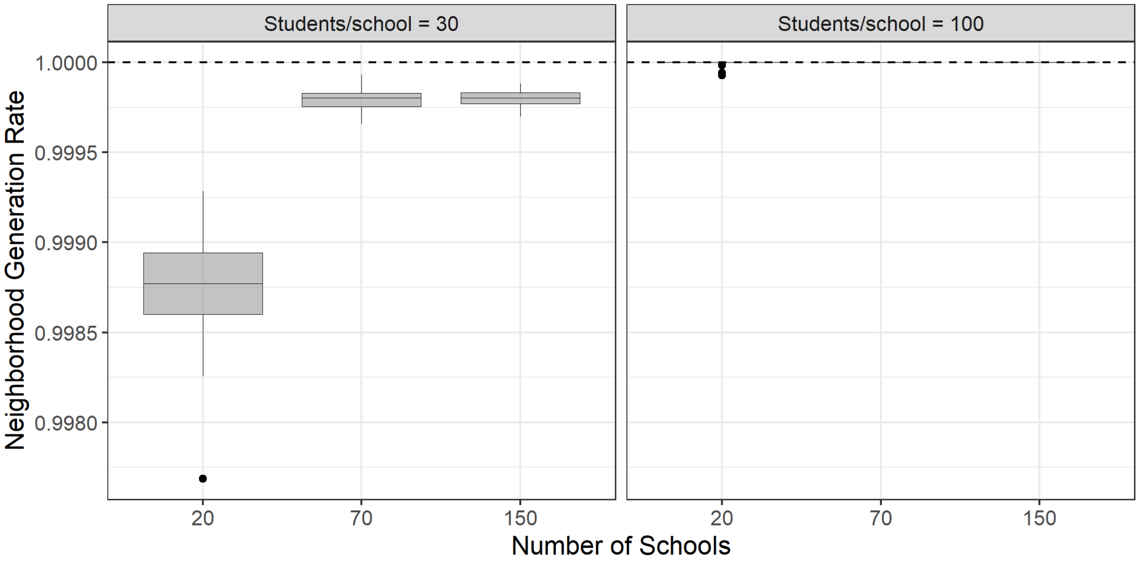

Figure 6.1. Neighborhood Generation Success Rates.

Note. Based on ANOVA analysis, the number of school clusters ($\eta_p^2 = 0.832$) and the number of students per school ($\eta_p^2 = 0.859$) had a large effect size on the success rate.

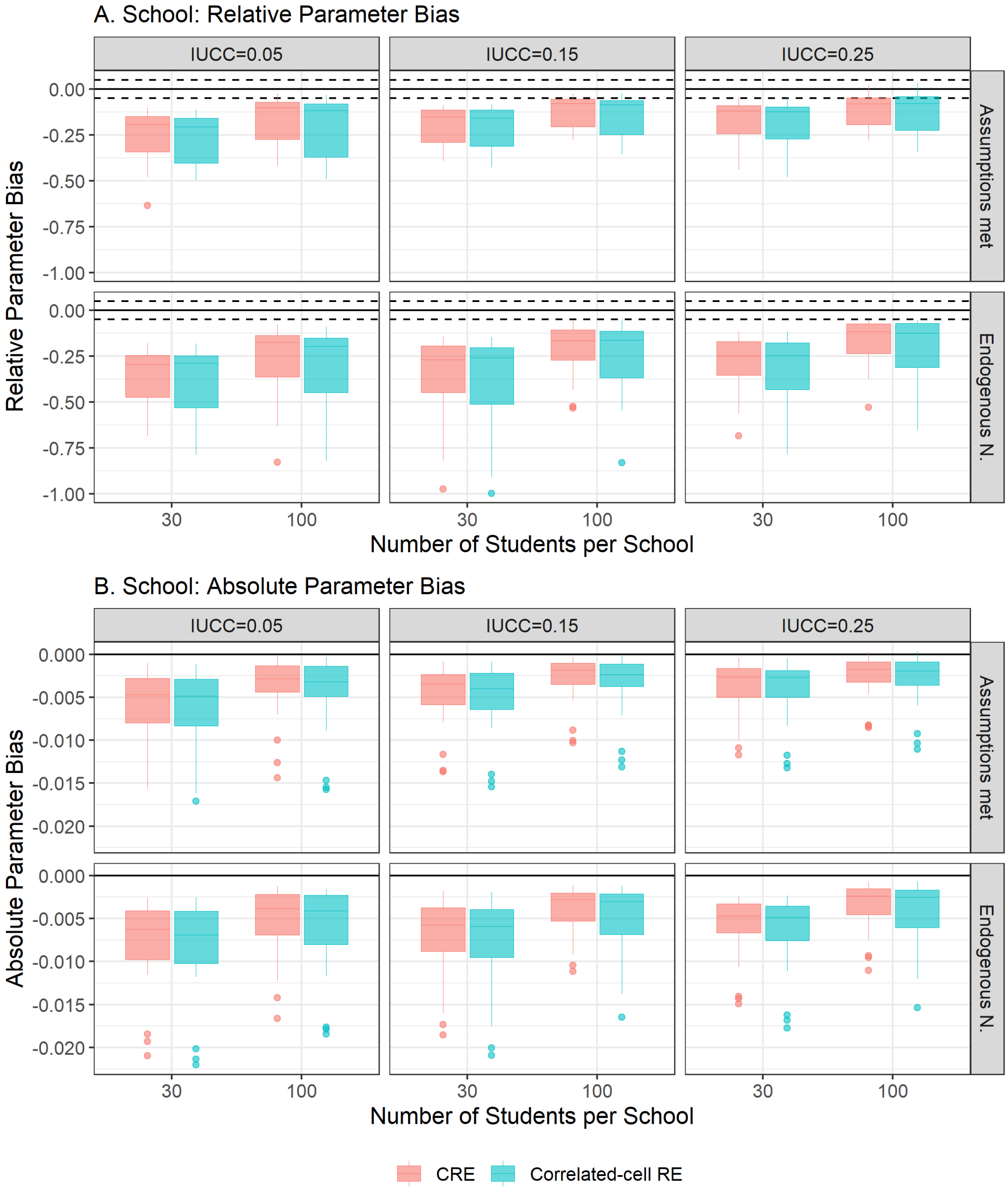

Figure 6.2. Relative and Absolute Parameter Bias by IUCC of the Between-School Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the school dimension was 0.049, while the maximum MCSE for absolute parameter bias was 0.001. IUCC in the columns indicates the neighborhood IUCC. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

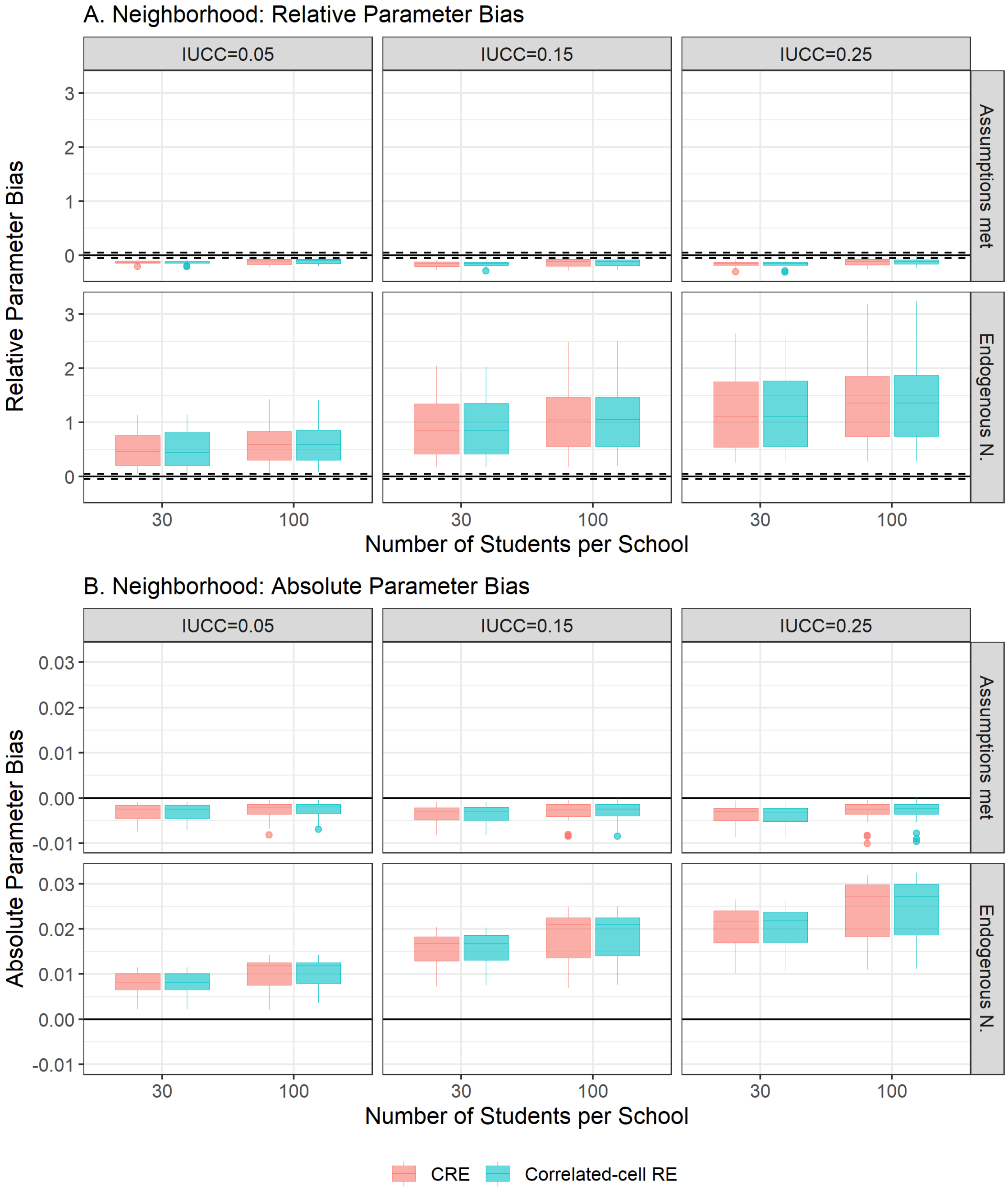

Figure 6.3. Relative and Absolute Parameter Bias by IUCC of the Between-Neighborhood Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the neighborhood dimension was 0.036, while the maximum MCSE for absolute parameter bias was 0.0004. IUCC in the columns indicates the neighborhood IUCC. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

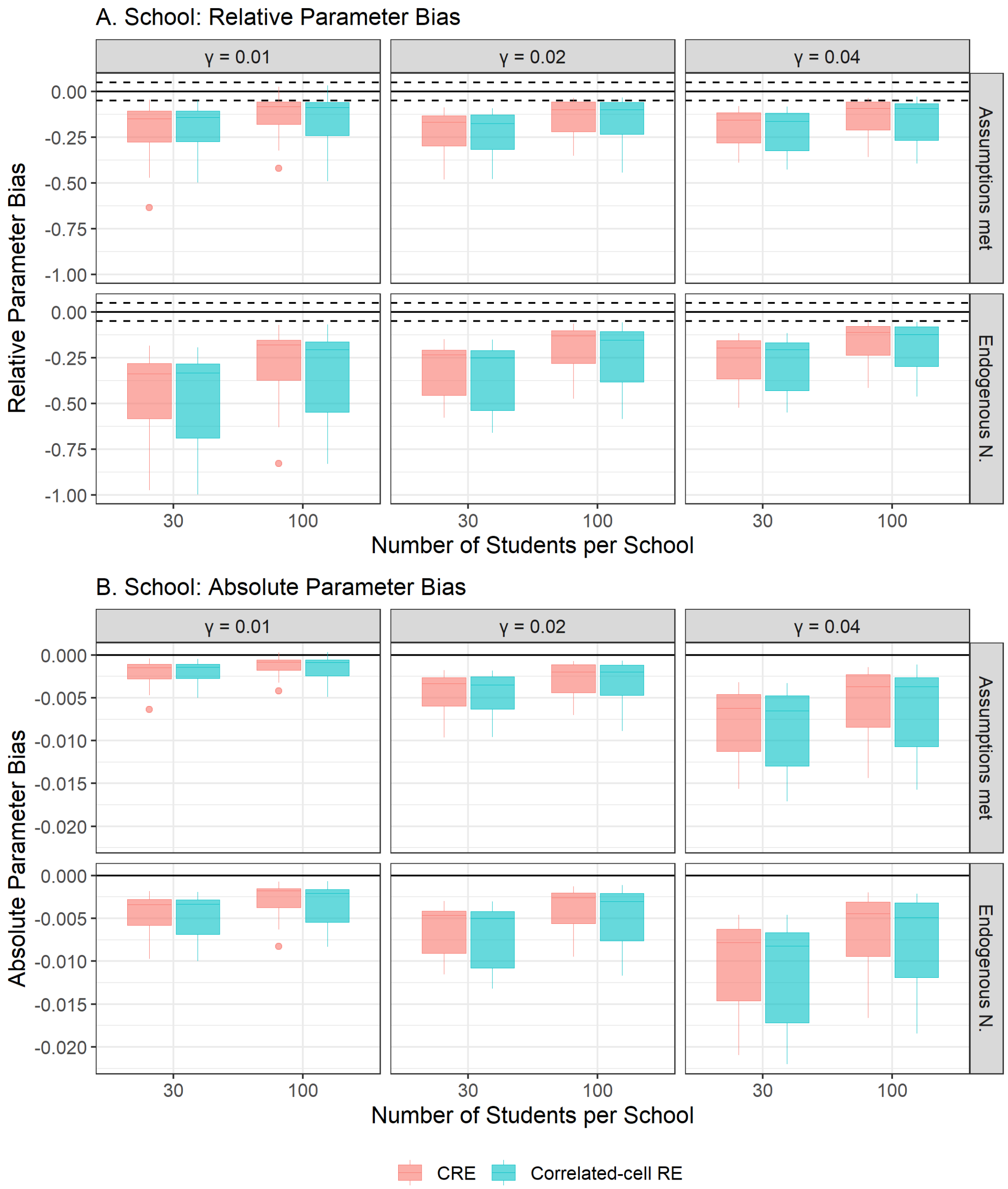

Figure 6.4. Relative and Absolute Parameter Bias by Coefficient Size of the Between-School Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the school dimension was 0.049, while the maximum MCSE for absolute parameter bias was 0.001. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

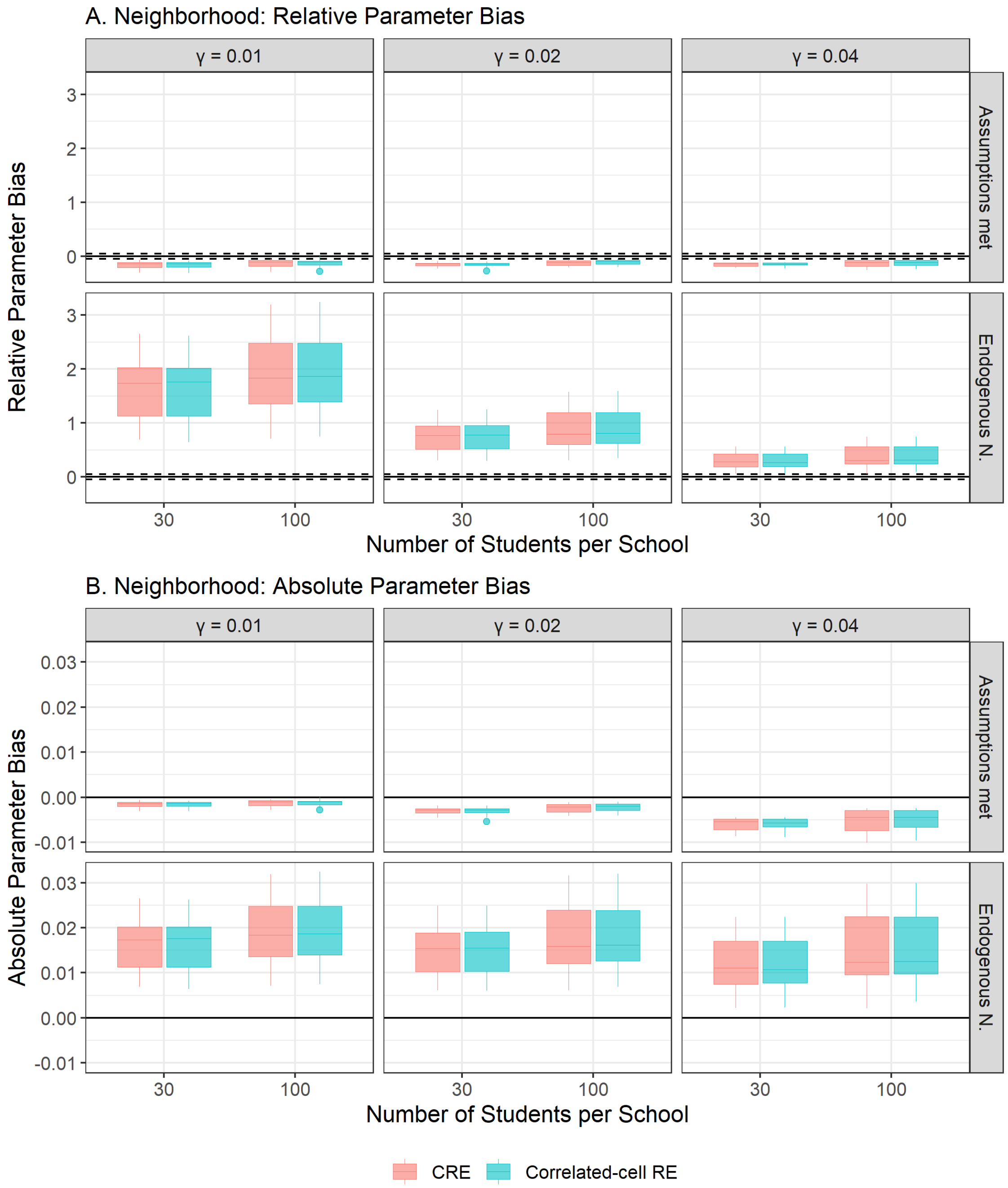

Figure 6.5. Relative and Absolute Parameter Bias by Coefficient Size of the Between-Neighborhood Effects for the Level-1 Covariate.

Note. The maximum MCSE for relative parameter bias in the neighborhood dimension was 0.036, while the maximum MCSE for absolute parameter bias was 0.0004. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

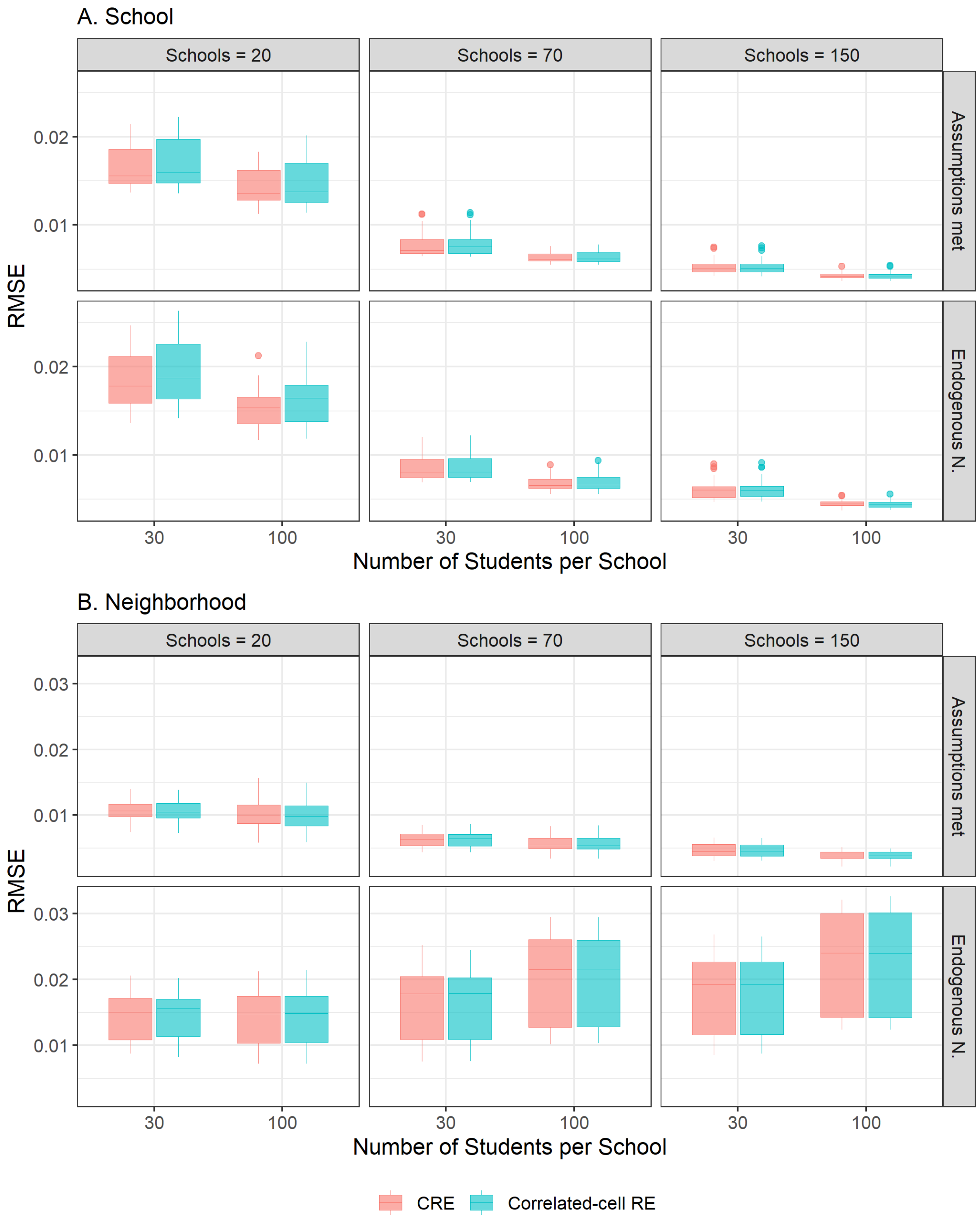

Figure 6.6. Root Mean Squared Error by the Number of Schools of the Between-Cluster Effects for the Level-1 Covariate.

Note. The maximum MCSE was 0.001 for the school dimension and 0.0004 for the neighborhood dimension. In the neighborhood dimension (B), the IUCC condition in columns indicates neighborhood IUCC. Endogenous N. indicates the exogeneity assumption was violated in neighborhood random effects.

6.3 The Number of Students per Cell

Though the number of students per cell was not included as one of the simulation conditions in this study, the number of students per cell can be calculated and included in ANOVA as a potential future simulation condition. The number of students per cell is calculated by the total number of students divided by the number of cells filled:

$$n_{cell} = \frac{J \times n_j}{0.1(J \times K)}, \tag{6.5}$$

where J is the number of schools, n_j is the number of students per school, 0.1 is the degree of sparsity, and $K = 3.5 \times J$ is the number of neighborhoods. Table 6.3 shows the calculated number of students per cell.

Table 6.3

Calculation of the Number of Students per Cell

Number of School	20	20	70	70	150	150
Number of Students per School	30	100	30	100	30	100
Number of Students per Cell	4	14	1	4	1	2

The results of incorporating the calculated number of students per cell into ANOVA are shown in Tables 6.4, 6.5, and 6.6. While the number of students per cell showed a negligible or small effect size for the parameter bias ($\eta_p^2 = 0.002$) and the relative bias of SE ($\eta_p^2 = 0.037$), respectively, the RMSE results revealed that the number of students per cell had a substantial effect size ($\eta_p^2 = 0.441$). Based on these ANOVA results, I plotted the relationship between the calculated number of students per cell and RMSE in Figure 6.7 to examine the impact of the number of students per cell on RMSE. In Figure 6.7, the number of level-1 students per cell did not exhibit a consistent direction of the effects on RMSE, indicating that additional simulation conditions should be considered along with the number of students per cell.

Table 6.4

ANOVA Results on Relative and Absolute Parameter Bias of the Within-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

	Relative PB	Absolute PB
Experimental Factors	η_p^2	η_p^2
Method	0.615 (large)	0.813 (large)
CCREM Assumption	0.425 (large)	0.675 (large)
Coefficient	0.184 (large)	0.001
Number of level-2 clusters (schools)	0.002	0.004
Number of level-1 students per school	0.084 (medium)	0.204 (large)
Neighborhood IUCC	0.002	0.008
Cell IUCC	0.004	0.014 (small)
Number of level-1 students per cell	0.002	0.002

Note. PB indicates parameter bias; Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Table 6.5

ANOVA Results on Root Mean Square Error of the Within-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

Experimental Factors	η_p^2
Method	0.764 (large)
CCREM Assumption	0.489 (large)
Coefficient	0.003
Number of level-2 clusters (schools)	0.948 (large)
Number of level-1 students per school	0.928 (large)
Neighborhood IUCC	0.194 (large)
Cell IUCC	0.002
Number of level-1 students per cell	0.441 (large)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

In Figure 6.8, I plotted the number of students per cell in conjunction with the number of schools and students per school. Figure 6.8 has some blank spots because the number of students per cell was not a pre-existing condition and thus not perfectly

Table 6.6

ANOVA Results on Relative Bias of Standard Error of the Within-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

Experimental Factors	η_p^2
Method	0.139 (medium)
CCREM Assumption	0.006
Coefficient	0.007
Number of level-2 clusters (schools)	0.190 (large)
Number of level-1 students per school	0.093 (medium)
Neighborhood IUCC	0.002
Cell IUCC	0.290 (large)
Number of level-1 students per cell	0.037 (small)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

matched with the other conditions. The graph shows that the number of schools and students per school had a more substantial influence than the number of students per cell. For example, when the number of students per school was constant, the RMSE was smaller (i.e., better performance) when the number of schools increased despite fewer students per cell. Also, when the number of schools remained constant and the number of students per school increased, the RMSE decreased with a greater number of students per cell. However, since the number of students per school influences the number of students per cell, it is challenging to determine the individual effect of students per cell. In summary, while the number of students per cell exhibited a substantial effect size, its magnitude remained smaller than the impact of the number of schools or students per school. Still, the effect of the number of students per cell should be investigated in future studies as an initially manipulated condition.

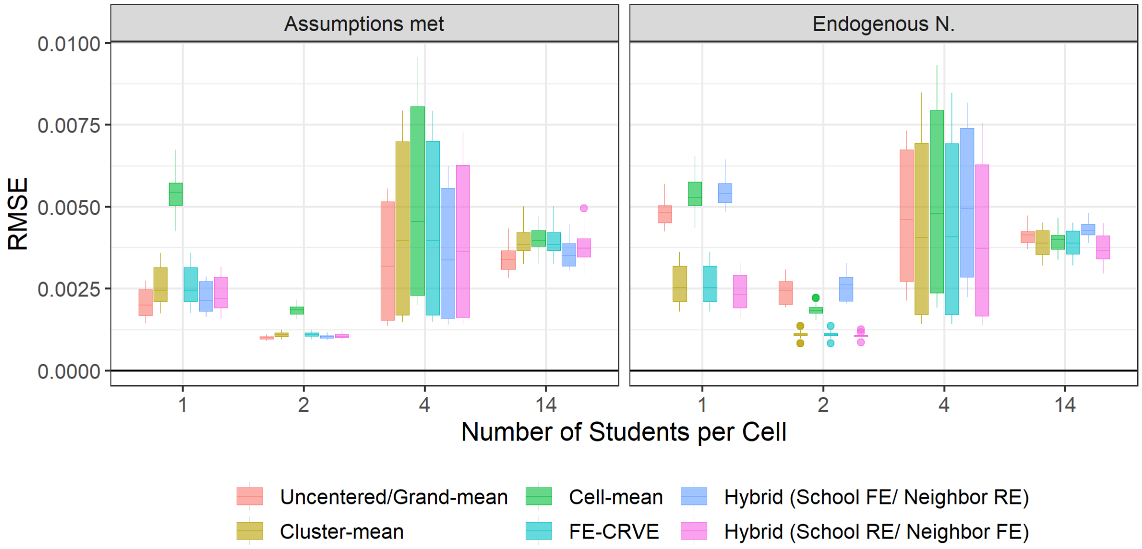

Figure 6.7. Root Mean Square Error of the Within-Cluster Effects for the Level-1 Covariate by the Number of Level-1 Students per Cell.

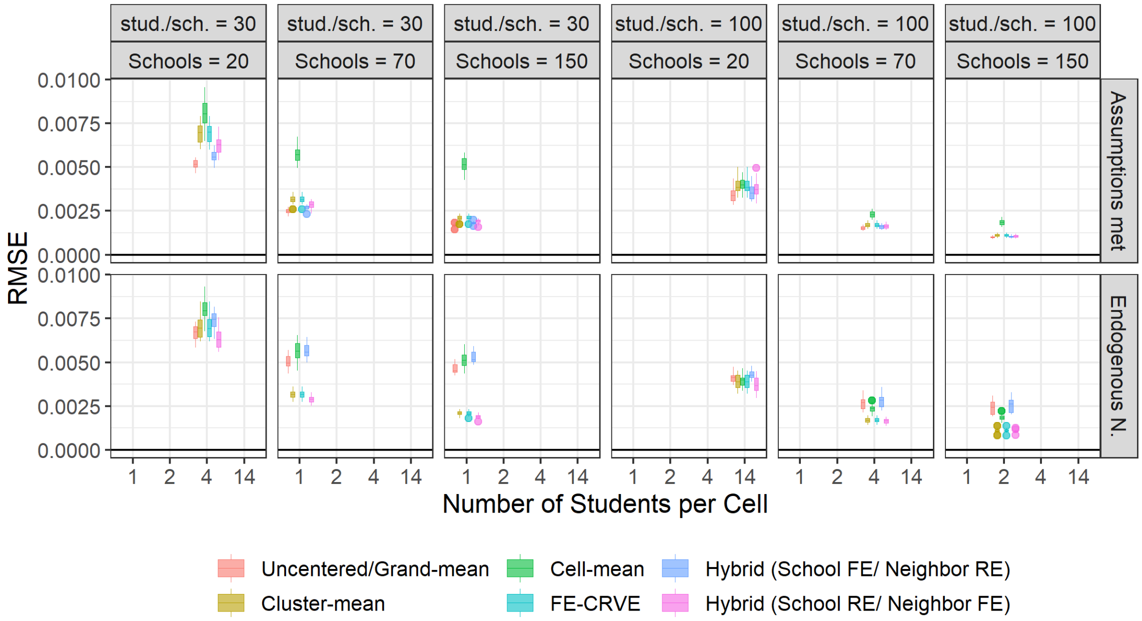

Figure 6.8. Root Mean Square Error of the Within-Cluster Effects for the Level-1 Covariate by the Number of Level-1 Students per Cell, the Number of Schools, and the Number of Level-1 Students per School.

Additionally, I presented the ANOVA results for the between-cluster effects, including the number of students per cell (see Tables 6.7). There is a slight variation, but the effect sizes of the number of students per cell were substantial in the parameter bias of the neighborhood clustering dimension and RMSE. The impact on the relative bias of SE was minimal. I omitted the graph for the between-cluster effect due to the combined impact of the number of students in the cell and other conditions.

Table 6.7

ANOVA Results on Relative and Absolute Parameter Bias of the Between-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

	School		Neighborhood	
Experimental Factors (η_p^2)	Relative PB	Absolute PB	Relative PB	Absolute PB
Method	0.075 (medium)	0.152 (large)	0.001	0.006
CCREM Assumption	0.627 (large)	0.676 (large)	0.961 (large)	0.995 (large)
Coefficient	0.317 (large)	0.912 (large)	0.873 (large)	0.853 (large)
Number of schools	0.897 (large)	0.947 (large)	0.570 (large)	0.925 (large)
Number of students/school	0.504 (large)	0.685 (large)	0.175 (large)	0.649 (large)
Neighborhood IUCC	0.324 (large)	0.532 (large)	0.653 (large)	0.932 (large)
Cell IUCC	0.008	0.009	0.003	0.007
Number of students/cell	0.009	0.018 (small)	0.063 (medium)	0.412 (large)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Table 6.8

ANOVA Results on Root Mean Squared Error of the Between-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

	School	Neighborhood
Experimental Factors	η_p^2	η_p^2
Method	0.052 (small)	0.000
CCREM Assumption	0.478 (large)	0.972 (large)
Coefficient	0.865 (large)	0.049 (small)
Number of level-2 clusters (schools)	0.989 (large)	0.009
Number of level-1 students per school	0.748 (large)	0.220 (large)
Neighborhood IUCC	0.346 (large)	0.922 (large)
Cell IUCC	0.168 (large)	0.040 (small)
Number of level-1 students per cell	0.151 (large)	0.250 (large)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

Table 6.9

ANOVA Results on Relative Bias of Standard Error of the Between-Cluster Effects for the Level-1 Covariate, including the Number of Level-1 Students per Cell

	School	Neighborhood
Experimental Factors	η_p^2	η_p^2
Method	0.003	0.006
CCREM Assumption	0.002	0.050 (small)
Coefficient	0.037 (small)	0.074 (medium)
Number of level-2 clusters (schools)	0.028 (small)	0.052 (small)
Number of level-1 students per school	0.015 (small)	0.013 (small)
Neighborhood IUCC	0.005	0.005
Cell IUCC	0.000	0.004
Number of level-1 students per school	0.003	0.018 (small)

Note. Small = .01, medium = .06, and large = .14; ANOVA was conducted in a two-way factorial, but only the main effects are shown in the table.

References

Abdel Magid, H. S., Milliren, C. E., Pettee Gabriel, K., & Nagata, J. M. (2021). Disentangling individual, school, and neighborhood effects on screen time among adolescents and young adults in the United States. *Preventive Medicine*, *142*, 106357. https://doi.org/10.1016/j.ypmed.2020.106357

Abdul-Hameed, B., & Matanmi, O. G. (2021). A Modified Breusch–Pagan Test for Detecting Heteroscedasticity in the Presence of Outliers. *Pure and Applied Mathematics Journal*, *10*(6), 139. https://doi.org/10.11648/j.pamj.20211006.13

Allensworth, E. M., & Luppescu, S. (2018). Why Do Students Get Good Grades, or Bad Ones? The Influence of The Teacher, Class, School, and Student. Working Paper. Retrieved January 21, 2023, from https://files.eric.ed.gov/fulltext/ED588781.pdf

Allison, P. D. (2005). *Fixed Effects Regression Methods for Longitudinal Data using SAS*. SAS Press.

Allison, P. (2009). *Fixed effects regression models*. Sage.

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On Ignoring the Random Effects Assumption in Multilevel Models: Review, Critique, and Recommendations. *Organizational Research Methods*, *24*(2), 443–483. https://doi.org/10.1177/1094428119877457

Asparouhov, T., & Muthén, B. (2019). Latent Variable Centering of Predictors and Mediators in Multilevel and Time-Series Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Attell, B. K. (2020). Changing Attitudes Toward Euthanasia and Suicide for Terminally Ill Persons, 1977 to 2016: An Age-Period-Cohort Analysis. *OMEGA*

- *Journal of Death and Dying*, *80*(3), 355–379. https://doi.org/10.1177/0030222817729612

Baird, J.-A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert- and Supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 44–59. https://doi.org/10.1080/0969594X.2015.1108283

Baldwin, S., & Fellingham, G. (2012). Bayesian Methods for the Analysis of Small Sample Multilevel Data With a Complex Variance Structure. *Psychological methods*, *18*. https://doi.org/10.1037/a0030642

Barker, K. M., Dunn, E. C., Richmond, T. K., Ahmed, S., Hawrilenko, M., & Evans, C. R. (2020). Cross-classified multilevel models (CCMM) in health research: A systematic review of published empirical studies and recommendations for best practices. *SSM - Population Health*, *12*, 100661. https://doi.org/10.1016/j.ssmph.2020.100661

Bartels, B. L. (2008). Beyond "fixed versus random effects": A framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data. *The Society for Political Methodology*, *9*, 1–43.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bauer, A. M., Jakupcak, M., Hawrilenko, M., Bechtel, J., Arao, R., & Fortney, J. C. (2021). Outcomes of a health informatics technology-supported behavioral activation training for care managers in a collaborative care program. *Families, Systems, & Health*, *39*(1), 89–100. https://doi.org/10.1037/fsh0000523

Bayer-Oglesby, L., Zumbrunn, A., Bachmann, N., & Team, o. b. o. t. S. (2022). Social inequalities, length of hospital stay for chronic conditions and the mediating role of comorbidity and discharge destination: A multilevel analysis of hospital administrative data linked to the population census in Switzerland. *PLOS ONE*, *17*(8), e0272265. https://doi.org/10.1371/journal.pone.0272265

Beck, A. N., Finch, B. K., Lin, S.-F., Hummer, R. A., & Masters, R. K. (2014). Racial disparities in self-rated health: Trends, explanatory factors, and the changing role of socio-demographics. *Social Science & Medicine*, *104*, 163–177. https://doi.org/10.1016/j.socscimed.2013.11.021

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, *3*(1), 133–153. https://doi.org/10.1017/psrm.2014.7

Beretvas, S. N., & Murphy, D. L. (2013). An Evaluation of Information Criteria Use for Correct Cross-Classified Random Effects Model Selection. *The Journal of Experimental Education*, *81*(4), 429–463. https://doi.org/10.1080/00220973.2012.745467

Bergé, L. (2018). *Efficient estimation of maximum likelihood models with multiple fixed-effects: The R package FENmlm* (tech. rep. No. 18-13) [DEM Discussion Paper Series]. Department of Economics at the University of Luxembourg. https://wwwen.uni.lu/content/download/110162/1299525/file/2018_13

Biesanz, J., Deeb-Sossa, N., Papadakis, A., Bollen, K., & Curran, P. (2004). The Role of Coding Time in Estimating and Interpreting Growth Curve Models. *Psychological methods*, *9*, 30–52. https://doi.org/10.1037/1082-989X.9.1.30

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, *47*(5), 1287–1294. https://doi.org/10.2307/1911963

Brunton-Smith, I., Sturgis, P., & Williams, J. (2012). Is Success in Obtaining Contact and Cooperation Correlated with the Magnitude of Interviewer Variance? *Public Opinion Quarterly*, *76*(2), 265–286. https://doi.org/10.1093/poq/nfr067

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. https://doi.org/https://doi.org/10.1002/sim.2673

Cafri, G., Hedeker, D., & Aarons, G. A. (2015). An Introduction and Integration of Cross-Classified, Multiple Membership, and Dynamic Group Random-Effects Models. *Psychological methods*, *20*(4), 407–421. https://doi.org/10.1037/met0000043

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics*, *29*(2), 238–249. https://doi.org/10.1198/jbes.2010.07136

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), 317–372. https://doi.org/10.3368/jhr.50.2.317

Castellaneta, F., & Gottschalg, O. (2016). Does ownership matter in private equity? The sources of variance in buyouts' performance. *Strategic Management Journal*, *37*(2), 330–348. https://doi.org/10.1002/smj.2336

Castellano, K. E., Rabe-Hesketh, S., & Skrondal, A. (2014). Composition, Context, and Endogeneity in School and Teacher Comparisons. *Journal of Educational and Behavioral Statistics*, *39*(5), 333–367. https://doi.org/10.3102/1076998614547576

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 [Publisher: American Psychological Association]. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Dağli, Ü. Y., & Jones, I. (2013). The Longitudinal Effects of Kindergarten Enrollment and Relative Age on Children's Academic Achievement. *Teachers College Record*, *115*(3), 1–40. https://doi.org/10.1177/016146811311500308

Darandari, E. (2004). *Robustness of Hierarchical Linear Model Parameter Estimates Under Violations of Second-Level Residual Homoskedasticity and Independence Assumptions* (Doctoral dissertation). Doctoral dissertation, Florida State University. Retrieved December 22, 2021, from https://www.proquest.com/docview/305182787?pq-origsite=gscholar&fromopenview=true

Davidson, R., & MacKinnon, J. G. (2004). *Econometric Theory and Methods* (Vol. 5). Oxford University Press.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. *Review of Educational Research*, *79*(1), 69–102. https://doi.org/10.3102/0034654308325581

D'Haese, S., Van Dyck, D., De Bourdeaudhuij, I., Deforche, B., & Cardon, G. (2014). The association between objective walkability, neighborhood socio-economic status, and physical activity in Belgian children. *International Journal of Behavioral Nutrition and Physical Activity*, *11*(1), 104. https://doi.org/10.1186/s12966-014-0104-1

Dronkers, J., Levels, M., & de Heus, M. (2014). Migrant pupils' scientific performance: The influence of educational system features of origin and destination countries. *Large-scale Assessments in Education*, *2*(1), 3. https://doi.org/10.1186/2196-0739-2-3

Dunn, E. C., Milliren, C. E., Evans, C. R., Subramanian, S. V., & Richmond, T. K. (2015). Disentangling the Relative Influence of Schools and Neighborhoods on Adolescents' Risk for Depressive Symptoms. *American Journal of Public Health*, *105*(4), 732–740. https://doi.org/10.2105/AJPH.2014.302374

Efron, B., & Stein, C. (1981). The Jackknife Estimate of Variance [Publisher: Institute of Mathematical Statistics]. *The Annals of Statistics*, *9*(3), 586–596. Retrieved February 23, 2023, from https://www.jstor.org/stable/2240822

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

Evans, S. C., & Fite, P. J. (2019). Dual Pathways from Reactive Aggression to Depressive Symptoms in Children: Further Examination of the Failure Model. *Journal of Abnormal Child Psychology*, *47*(1), 85–97. https://doi.org/10.1007/s10802-018-0426-6

Feaster, D., Brincks, A., Robbins, M., & Szapocznik, J. (2011). Multilevel Models to Identify Contextual Effects on Individual Group Member Outcomes: A Family Example. *Family Process*, *50*(2), 167–183. https://doi.org/https://doi.org/10.1111/j.1545-5300.2011.01353.x

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279. https://doi.org/10.1080/02602938.2016.1261083

Feistauer, D., & Richter, T. (2018). The Role of Clarity About Study Programme Contents and Interest in Student Evaluations of Teaching. *Psychology Learning & Teaching*, *17*(3), 272–292. https://doi.org/10.1177/1475725718779727

Fielding, A., & Goldstein, H. (2006). *Cross-classified and Multiple Membership Structures in Multilevel Models: An introduction and review* (tech. rep. No. 791). DfES, London, UK.

Fleischmann, M., Hübner, N., Marsh, H. W., Guo, J., Trautwein, U., & Nagengast, B. (2022). Which class matters? Juxtaposing multiple class environments as frames-of-reference for academic self-concept formation. *Journal of Educational Psychology*, *114*(1), 127–143. https://doi.org/10.1037/edu0000491

Francis, D. J., Kulesz, P. A., & Benoit, J. S. (2018). Extending the Simple View of Reading to Account for Variation Within Readers and Across Texts: The Complete View of Reading (CVRi). *Remedial and Special Education*, *39*(5), 274–288. https://doi.org/10.1177/0741932518772904

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*, *28*(2), 221–239. https://doi.org/10.1002/sim.3478

Gaure, S. (2013a). Lfe: Linear group fixed effects. *The R Journal*, *5*(2), 114–117.

Gaure, S. (2013b). OLS with multiple high dimensional category variables. *Computational Statistics & Data Analysis*, *66*, 8–18. https://doi.org/10.1016/j.csda.2013.03.024

Gaure, S. (2014). Correlation bias correction in two-way fixed-effects linear regression. *Stat*, *3*(1), 379–390. https://doi.org/10.1002/sta4.68

Gilbert, J., Petscher, Y., Compton, D. L., & Schatschneider, C. (2016). Consequences of Misspecifying Levels of Variance in Cross-Classified Longitudinal Data Structures. *Frontiers in Psychology*, *7*.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, *73*, 43–56.

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). John Wiley & Sons.

Goldstein, H., & Sammons, P. (1997). The Influence of Secondary and Junior Schools on Sixteen Year Examination Performance: A Cross-classified Multilevel Analysis. *School Effectiveness and School Improvement*, *8*(2), 219–230. https://doi.org/10.1080/0924345970080203

Goodale, B. M., Shilaih, M., Falco, L., Dammeier, F., Hamvas, G., & Leeners, B. (2019). Wearable Sensors Reveal Menses-Driven Changes in Physiology and Enable Prediction of the Fertile Window: Observational Study. *Journal of Medical Internet Research*, *21*(4), e13404. https://doi.org/10.2196/13404

Greene, W. H. (2018). *Econometric Analysis* (8th ed.). Pearson.

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences*, *7*(4), 121–133.

Groenewegen, P. P., Zock, J.-P., Spreeuwenberg, P., Helbich, M., Hoek, G., Ruijsbroek, A., Strak, M., Verheij, R., Volker, B., Waverijn, G., & Dijst, M. (2018). Neighbourhood social and physical environment and general practitioner assessed morbidity. *Health & Place*, *49*, 68–84. https://doi.org/10.1016/j.healthplace.2017.11.006

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i02

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251–1271. https://doi.org/10.2307/1913827

Hayward, R. D., & Krause, N. (2015). Aging, Social Developmental, and Cultural Factors in Changing Patterns of Religious Involvement Over a 32-Year Period: An Age–Period–Cohort Analysis of 80 Countries. *Journal of Cross-Cultural Psychology*, *46*(8), 979–995. https://doi.org/10.1177/0022022115597066

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2008). An Application of a Mixed-Effects Location Scale Model for Analysis of Ecological Momentary Assessment (EMA) Data. *Biometrics*, *64*(2), 627–634. https://doi.org/10.1111/j.1541-0420.2007.00924.x

Hehman, E., & Sutherland, C. A. M. (2017). The Unique Contributions of Perceiver and Target Characteristics in Person Perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529.

Hoffman, L. (2019). On the Interpretation of Parameters in Multivariate Multilevel Models Across Different Combinations of Model Specification and Estimation. *Advances in methods and practices in psychological science*, *2*(3), 288–311. https://doi.org/10.1177/2515245919842770

Hofmann, D. A., & Gavin, M. B. (1998). Centering Decisions in Hierarchical Linear Models: Implications for Research in Organizations. *Journal of Management*, *24*(5), 623–641. https://doi.org/10.1177/014920639802400504

Hoogland, J. J., & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling: An Overview and a Meta-Analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hox, J. J., & Roberts, J. K. (Eds.). (2011). *Handbook of advanced multilevel analysis* [OCLC: 705730688]. Routledge.

Kendler, K. S., Ohlsson, H., Sundquist, K., & Sundquist, J. (2015). Environmental clustering of drug abuse in households and communities: Multi-level modeling of a national Swedish sample. *Social Psychiatry and Psychiatric Epidemiology*, *50*(8), 1277–1284. https://doi.org/10.1007/s00127-015-1030-5

Kennedy, P. (2008). *A Guide to Econometrics* (6th ed.). Wiley & Sons.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558

Kim, B., Spohn, C., & Hedberg, E. C. (2015). Federal Sentencing as a Complex Collaborative Process: Judges, Prosecutors, Judge–Prosecutor Dyads, and Disparity in Sentencing. *Criminology*, *53*(4), 597–623. https://doi.org/10.1111/1745-9125.12090

Kim, Y. K., & Sax, L. J. (2014). The Effects of Student–Faculty Interaction on Academic Self-Concept: Does Academic Major Matter? *Research in Higher Education*, *55*(8), 780–809. https://doi.org/10.1007/s11162-014-9335-x

Kreft, I. G., de Leeuw, J., & Aiken, L. S. (1995). The Effect of Different Forms of Centering in Hierarchical Linear Models. *Multivariate Behavioral Research*, *30*(1), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

LeBeau, B. (2013). *Misspecification of the covariance matrix in the linear mixed model: A Monte Carlo simulation* (Doctoral dissertation). Doctoral dissertation, The University of Minnesota. St. Paul, MN. Retrieved December 22, 2021, from https://www.proquest.com/docview/1322973438?pq-origsite=gscholar&fromopenview=true

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling Heterogeneous Variance–Covariance Components in Two-Level Models. *Journal of Educational and Behavioral Statistics*, *39*(5), 307–332. https://doi.org/10.3102/1076998614546494

Lee, Y. R., & Pustejovsky, J. E. (2023). Comparing random effects models, ordinary least squares, or fixed effects with cluster robust standard errors for cross-classified data. *Psychological Methods*. https://doi.org/10.1037/met0000538

Lee, Y., & Nelder, J. A. (2006). Double hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *55*(2), 139–185. https://doi.org/10.1111/j.1467-9876.2006.00538.x

Lei, X., Li, H., & Leroux, A. J. (2018). Does a teacher's classroom observation rating vary across multiple classrooms? *Educational Assessment, Evaluation and Accountability*, *30*(1), 27–46. https://doi.org/10.1007/s11092-017-9269-x

Lin, S.-F., Beck, A. N., & Finch, B. K. (2014). Black–White Disparity in Disability Among U.S. Older Adults: Age, Period, and Cohort Trends. *The Journals of*

Gerontology: Series B, *69*(5), 784–797. https://doi.org/10.1093/geronb/gbu010

Lin, S.-F., Beck, A. N., & Finch, B. K. (2016). The Dynamic contribution of chronic conditions to temporal trends in disability among U.S. adults. *Disability and Health Journal*, *9*(2), 332–340. https://doi.org/10.1016/j.dhjo.2015.11.006

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*(4), 817–827.

Lovell, M. C. (2008). A Simple Proof of the FWL Theorem. *Journal of Economic Education*, *39*(1), 88–91. https://doi.org/10.3200/JECE.39.1.88-91

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Luo, W., & Kwok, O. (2009). The Impacts of Ignoring a Crossed Factor in Analyzing Cross-Classified Data. *Multivariate Behavioral Research*, *44*(2), 182–212. https://doi.org/10.1080/00273170902794214

Luo, W., & Kwok, O. (2012). The Consequences of Ignoring Individuals' Mobility in Multilevel Growth Models: A Monte Carlo Study. *Journal of Educational and Behavioral Statistics*, *37*(1), 31–56. https://doi.org/10.3102/1076998610394366

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting Practice in Multilevel Modeling: A Revisit After 10 Years. *Review of Educational Research*, *91*(3), 311–355. https://doi.org/10.3102/0034654321991229

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Maeda, Y. (2007). *Monte Carlo evidence regarding the effects of violating assumed conditions of two-level hierarchical models for cross-sectional data* (Doctoral dissertation). Doctoral dissertation, The University of Minnesota. Retrieved December 22, 2021, from https://www.proquest.com/docview/304840748?pq-origsite=gscholar&fromopenview=true

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual Analysis through the Multilevel Linear Model. *Sociological Methodology*, *14*, 72–103. https://doi.org/10.2307/270903

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science*, *26*(3), 388–402. Retrieved April 13, 2023, from https://www.jstor.org/stable/23059138

McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods*, *24*(1), 20–35. https://doi.org/10.1037/met0000182

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. *Multivariate Behavioral Research*, *51*(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

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

Meyers, J. L., & Beretvas, S. N. (2006). The Impact of Inappropriate Modeling of Cross-Classified Data Structures. *Multivariate Behavioral Research*, *41*(4), 473–497. https://doi.org/10.1207/s15327906mbr4104_3

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/https://doi.org/10.1002/sim.8086

Morton, K. L., Corder, K., Suhrcke, M., Harrison, F., Jones, A. P., van Sluijs, E. M. F., & Atkin, A. J. (2016). School polices, programmes and facilities, and objectively measured sedentary time, LPA and MVPA: Associations in secondary school and over the transition from primary to secondary school. *International Journal of Behavioral Nutrition and Physical Activity*, *13*(1), 54. https://doi.org/10.1186/s12966-016-0378-6

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica*, *46*(1), 69–85. https://doi.org/10.2307/1913646

Nisic, N., & Melzer, S. M. (2016). Explaining Gender Inequalities That Follow Couple Migration. *Journal of Marriage and Family*, *78*(4), 1063–1082. https://doi.org/10.1111/jomf.12323

Nuño, L. E., & Katz, C. M. (2019). Understanding Gang Joining from a Cross Classified Multi-Level Perspective. *Deviant Behavior*, *40*(3), 301–325. https://doi.org/10.1080/01639625.2017.1421706

O'Brien, R. M. (1990). Estimating the Reliability of Aggregate-Level Variables Based on Individual-Level Characteristics. *Sociological Methods & Research*, *18*(4), 473–504. https://doi.org/10.1177/0049124190018004004

Paccagnella, O. (2006). Centering or Not Centering in Multilevel Models? The Role of the Group Mean and the Assessment of Group Effects. *Evaluation Review*, *30*(1), 66–85. https://doi.org/10.1177/0193841X05275649

Pae, H. K., Bae, S., & Yi, K. (2020). Lexical properties influencing visual word recognition in Hangul. *Reading and Writing*, *33*(9), 2391–2412. https://doi.org/10.1007/s11145-020-10042-4

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, n160. https://doi.org/10.1136/bmj.n160

Palta, M., & Seplaki, C. (2002). Causes, Problems and Benefits of Different Between and Within Effects in the Analysis of Clustered Data. *Health Services and Outcomes Research Methodology*, *3*(3), 177–193.

Park, H.-C., Kim, D.-K., Kho, S.-Y., & Park, P. Y. (2017). Cross-classified multilevel models for severity of commercial motor vehicle crashes considering heterogeneity among companies and regions. *Accident Analysis & Prevention*, *106*, 305–314. https://doi.org/10.1016/j.aap.2017.06.009

Patchan, M. M., Schunn, C. D., & Correnti, R. J. (2016). The nature of feedback: How peer feedback features affect students' implementation rate and quality of revisions. *Journal of Educational Psychology*, *108*(8), 1098–1120. https://doi.org/10.1037/edu0000103

Paterson, L. (1991). Socio-economic status and educational attainment: A multidimensional and multi-level study. *Evaluation & Research in Education*, *5*(3), 97–121. https://doi.org/10.1080/09500799109533303

Patton, S. A. (2019). *Embedding Explicit Instruction of Transfer to Improve At-Risk Students' Reading Comprehension in Informational Texts* (Ph.D.) [ISBN: 9798582577614]. Vanderbilt University. United States – Tennessee. Retrieved January 12, 2023, from https://www.proquest.com/docview/2507668881/abstract/E33FF078D51844EAPQ/1

Pedersen, W., Bakken, A., & von Soest, T. (2018). Neighborhood or School? Influences on Alcohol Consumption and Heavy Episodic Drinking Among Urban Adolescents. *Journal of Youth and Adolescence*, *47*(10), 2073–2087. https://doi.org/10.1007/s10964-017-0787-0

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies*, *22*(1), 435–480. https://doi.org/10.1093/rfs/hhn053

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

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rast, P., Hofer, S. M., & Sparks, C. (2012). Modeling Individual Differences in Within-Person Variation of Negative and Positive Affect in a Mixed Effects Location Scale Model Using BUGS/JAGS. *Multivariate Behavioral Research*, *47*(2), 177–200. https://doi.org/10.1080/00273171.2012.658328

Raudenbush, S. W. (1989). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter*, *1*(2), 10–12.

Raudenbush, S. W. (2009). Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings. *Education Finance and Policy*, *4*(4), 468–491. https://doi.org/10.1162/edfp.2009.4.4.468

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

Raudenbush, S. W., & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education*, *59*(1), 1–17. https://doi.org/10.2307/2112482

Raudenbush, S. W. (1993). A Crossed Random Effects Model for Unbalanced Data With Applications in Cross-Sectional and Longitudinal Research. *Journal of Educational Statistics*, *18*(4), 321–349. https://doi.org/10.3102/10769986018004321

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A Multilevel, Multivariate Model for Studying School Climate With Estimation Via the EM Algorithm and Application to U.S. High-School Data. *Journal of Educational Statistics*, *16*(4), 295–330. https://doi.org/10.3102/10769986016004295

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*, *22*(2), 154–184. https://doi.org/10.1080/15305058.2022.2036161

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/https://doi.org/10.1111/2041-210X.13434

Schmidt, J. A., O'Neill, T. A., & Dunlop, P. D. (2021). The Effects of Team Context on Peer Ratings of Task and Citizenship Performance. *Journal of Business and Psychology*, *36*(4), 573–588. https://doi.org/10.1007/s10869-020-09701-8

Sharp, G., Denney, J. T., & Kimbro, R. T. (2015). Multiple contexts of exposure: Activity spaces, residential neighborhoods, and self-rated health. *Social Science & Medicine*, *146*, 204–213. https://doi.org/10.1016/j.socscimed.2015.10.040

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 1–15. https://doi.org/10.1348/000711008X398968

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. https://doi.org/10.3102/1076998609345252

Silber, H., Roßmann, J., Gummer, T., Zins, S., & Weyandt, K. W. (2021). The effects of question, respondent and interviewer characteristics on two types of item nonresponse. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(3), 1052–1069. https://doi.org/10.1111/rssa.12703

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* (2nd ed.). Sage.

Stan Development Team. (2023). *RStan: The R interface to Stan.* R package version 2.21.8. https://mc-stan.org/

Stevens, J. P. (2007). *Intermediate Statistics: A Modern Approach.* (3rd ed.). Routledge.

Thompson, S. B. (2011). Simple formulas for standard errors that cluster by both firm and time. *Journal of Financial Economics*, *99*(1), 1–10. https://doi.org/10.1016/j.jfineco.2010.08.016

Thornton III, G. C., Rupp, D. E., Gibbons, A. M., & Vanhove, A. J. (2019). Same-gender and same-race bias in assessment center ratings: A rating error approach to understanding subgroup differences. *International Journal of Selection and Assessment*, *27*(1), 54–71. https://doi.org/10.1111/ijsa.12229

Thrash, T. M., Maruskin, L. A., Moldovan, E. G., Oleynick, V. C., & Belzak, W. C. (2017). Writer–reader contagion of inspiration and related states: Conditional process analyses within a cross-classified writer × reader framework. *Journal of Personality and Social Psychology*, *113*(3), 466–491. https://doi.org/10.1037/pspp0000094

Vagi, R. L., Collins, C., & Clark, T. (2017). Identifying scalable policy solutions: A state-wide cross-classified analysis of factors related to early childhood literacy. *Education Policy Analysis Archives*, *25*, 9. https://doi.org/10.14507/epaa.25.2686

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-Classification Multilevel Logistic Models in Psychometrics. *Journal of Educational and Behavioral Statistics*, *28*(4), 369–386. https://doi.org/10.3102/10769986028004369

van Berkel, S. R., Groeneveld, M. G., van der Pol, L. D., Linting, M., & Mesman, J. (2022). Growing up together: Differences between siblings in the development of compliance separating within-family and between-family effects. *Developmental Psychology*. https://doi.org/10.1037/dev0001486

van Braak, M., van de Pol, J., Poorthuis, A. M. G., & Mainhard, T. (2021). A microperspective on students' behavioral engagement in the context of teachers' instructional support during seatwork: Sources of variability and the role of teacher adaptive support. *Contemporary Educational Psychology*, *64*, 101928. https://doi.org/10.1016/j.cedpsych.2020.101928

Weiser, B. L. (2013). Ameliorating Reading Disabilities Early: Examining an Effective Encoding and Decoding Prevention Instruction Model. *Learning Disability Quarterly*, *36*(3), 161–177. https://doi.org/10.1177/0731948712450017

White, H. (1984). *Asymptotic theory for econometricians.* Academic press.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data.* MIT Press.

Wooldridge, J. M. (2003). *Introductory Econometrics: A Modern Approach.* South-Western College Publishing.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data.* (2nd ed.). MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Cengage Learning.

Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender [Publisher: American Psychological Association]. *Journal of Personality and Social Psychology*, *117*(2), 364–385. https://doi.org/10.1037/pspi0000160

Yang, Y., & Land, K. C. (2008). Age–Period–Cohort Analysis of Repeated Cross-Section Surveys: Fixed or Random Effects? *Sociological Methods & Research*, *36*(3), 297–326. https://doi.org/10.1177/0049124106292360

Zhang, L. (2017). An Age–Period–Cohort Analysis of Religious Involvement and Adult Self-Rated Health: Results from the USA, 1972–2008. *Journal of Religion and Health*, *56*(3), 916–945. https://doi.org/10.1007/s10943-016-0292-x