QUANTIFYING THE BIAS OF STANDARD ERROR ESTIMATES DUE TO OMITTED CLUSTER LEVELS IN COMPLEX MULTILEVEL DATA: A SENSITIVITY ANALYSIS FOR EMPIRICAL RESEARCHERS

By

Zixi Chen

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Measurement and Quantitative Methods—Doctor of Philosophy

2020

ABSTRACT

QUANTIFYING THE BIAS OF STANDARD ERROR ESTIMATES DUE TO OMITTED CLUSTER LEVELS IN COMPLEX MULTILEVEL DATA: A SENSITIVITY ANALYSIS FOR EMPIRICAL RESEARCHERS

By

Zixi Chen

Educational phenomena occur in multilevel contexts, such as students nested within classrooms and classrooms nested within schools. This multilevel structure is also reflected in the multi-stage sampling design and randomized experimental design by clusters in educational data collection and research design. The consequential challenge of dependent observations within clusters of each nesting layer is prevalently dealt with by Hierarchical Linear Modeling (HLM) in education studies. However, in many cases, the observed data's multilevel structure can be unidentified or misspecified that the complex multilevel data structure is partially presented. Thus, even with the advanced statistical tools, the estimated models with omitted clustering levels will still produce erroneous standard error estimates and result in either Type I or Type II errors that compromise and even undistort interpretations of educational mechanisms. Practical guidance is urgently needed for empirical research confronting this issue to judge and detect whether the estimated models are adequate in taking account of the clustering dependency.

This paper contributes to investigate when a cluster level should be explicitly modeled but omitted and how much the standard error estimates would be biased. This paper examines these questions in settings of a true three-level clustered data structure, while a cluster level, either at the highest, middle, or the lowest level, is omitted in the estimated two-level models. The theoretical discussion of essential clustering levels in modeling due to multi-stage sampling design and randomized experiments by clusters is drawn on insights from Abadie et al. (2017)

and Hedges and Rhoads (2011). The current study then derives corresponding mathematical formulas to quantify the standard error estimation bias for each level's predictors' estimated effect. These derived formulas are functions of the intraclass correlation coefficients and cluster sizes of the estimated and omitted cluster levels. Further, build on Frank, Maroulis, Duong, and Kelcey (2013), the current paper develops a sensitivity analysis framework with a scientific language to quantify the degree of statistical inferences robustness based on the clustering characteristics of the omitted levels of clusters. In each omitted clustering scenario at the lowest, middle, and highest level, empirical studies are provided as implication examples of the sensitivity analysis to demonstrate the potential inference robustness risks due to omitted clustering.

ACKNOWLEDGEMENTS

I would like to express the heartfelt gratitude to my advisor and committee Chair Dr. Kenneth Frank for giving me invaluable guidance throughout my doctoral study and this research. His vision, experience, and sincerity in research have deeply inspired me. He taught me the spirit to grow to be a good scientist and the philosophy to conduct scientific research. His kindness and empathy have continuously supported and encouraged me to overcome all the challenges an international student faces. Without his support, my doctoral study would not have been possible.

I would also like to extend my sincere appreciations to my committee, Dr. Jeffery Wooldridge, Dr. Kimberly Kelly, and Dr. Spyros Konstantopoulos for their expertise and support. Their insights on this study has largely inspired me to learn the different languages of economics, education, and statistics while seeking for the common ground and the essence of the research goals. I would like to acknowledge the Dissertation Completion Fellowship from the Graduate School of Michigan State University financially supported this study.

I owe my deepest gratitude to my parents, Zaimin Chen and Yueping Yu, and my grandparents, who love me and always support me unconditionally. For the past nine years, I have been living in another continent; my family's endless love is what carries me to accomplish this journey. I would also like to thank the long-standing support and love from my fiancé, Haoran Tan, and his family.

Last but not the least, I sincerely appreciate my dear friends and colleagues for their intellectual sharing, emotional support, and companions: Dr. Kaitlin Knake, Dr. Qinyun Lin,

Yuqing Liu, Jordan Tait, Dr. Kimberly Jensen, Dr. Tenglong Li, Dr. Jihyun Kim, Dr. Sihua Hu, Dr. Faran Zhou, Dr. Zheng Gu, Wanqing Apa, Bixi Zhang, and Xuran Wang. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Educational phenomena occur in a nested context, such as students nested within classes within schools (Barr & Dreeben, 1983). In this multilevel schooling system, higher-level school actors, such as administrators and principals, as well as school social context, shape and respond to educational activities of the lower-level actors, such as teachers and students, through flows of resource allocations and routine organizational designs (Gamoran & Dreeben,1986; Goddard et al., 2007; Hallinger, & Murphy, 1986; Heck et al., 1990; Seashore Louis & Lee, 2016; Spillane et al., 2011). These units, such as schools, classrooms, and students, are inherent (or innate) levels in the formed organizational system of schooling (Krull & MacKinnon, 2001).

The multi-stage sampling design of education data collection corresponds to the multilevel structure of schooling system (Konstantopoulos, 2008a, 2008b; Snijders & Bosker, 2011; Hedges & Rhoads, 2011). Larger units, such as schools from the population of interest, are randomly selected in the first stage and are referred to as Primary Sampling Units (PSUs) (Leeuw & Meijer, 2008). In the second stage, researchers sample smaller units, such as classrooms, from PSUs. The sampled students are thus Secondary Sampling Units (SSUs), which are nested within school clusters. Stages of sampling can continue until the Ultimate Sampling Units (USUs), which are normally the targeted research units, such as students, are reached (Battaglia,2008). These sampling stages define the deliberate cluster levels by design in analysis.

The multilevel nesting structure results in dependencies between individual actors within clusters, challenging the independent observation assumption of the conventional regression analysis using Ordinary Least Squares (OLS) estimation. For instance, students who are similar in motivation, achievement, and family background are more likely to be grouped in the same classrooms and schools (Goldstein, 2011; Snijders & Bosker, 2011). It is also possible that students become more similar after they are assigned to the same classes and schools, as they share similar learning experiences and social contexts (see empirical examples in Frank, Muller, et al, 2008, and Rhoads, 2011). Teachers could become similar in instructions through professional training, collaboration and social interactions which will ultimately expose to their students learning activities, form within-school shared culture, and collectively react to policy enactment (see empirical examples in Coburn et al., 2012; Goddard et al., 2007; Penuel et al., 2009, and a survey in Voogt et al., 2016). With the clustering dependency, the independent error assumption treating the data as a simple random sample is thus violated. It is well-documented that the standard error estimates of coefficients from OLS estimation are underestimated though the coefficient estimates are unbiased, which leads to Type I errors (McNeish, 2014; Mundfrom & Schults, 2002; Musca et al., 2011).

The research interest of multilevel-structured educational and social phenomena and the methodological needs of dealing with the clustering dependency lead to the prevalent use of Hierarchical Linear Model (HLM) (Raudenbush & Bryk, 2002; Frank, 1998; Musca et al., 2011; Niehaus et al., 2014; Snijders & Bosker, 2011). HLM explicitly models the multilevel clustering dependency by including the corresponding level's random effects that capture the between-cluster variation and identify the cluster-specific effects beyond the population-averaged estimates (McNeish et al., 2017; Snijders & Bosker, 2011). In a three-stage sampled data

structure, a three-level HLM model can account for dependencies of USUs nested within SUSs within PSUs. This structure aligns well with the conventional schooling system mentioned above, that students are nested within classes, and classes are nested within schools.

The advantages of using HLM to make robust statistical inferences with clustered data could easily vanish if the imperative modeling assumptions relevant to random effects do not hold true (Dedrick, et al., 2009; Huang, 2018; McNeish et al., 2017; Snijders & Berkhof, 2008). Since the random effects variance is taken into account in estimating the standard errors of the regression coefficients (Raudenbush & Bryk, 2002; Snijders & Bosker, 2011), an essential assumption is that the modeled cluster levels as random effects are sufficient and the corresponding random effects are correctly specified. For example, in practice, researchers may purposively exclude a cluster level, such as the classrooms, in modeling for parsimony, regardless of testing whether this omission would result in ignoring clustering dependency and false inferences.

1.2 Problem Statement

The omission of cluster levels or misspecified random effects masks some true sources of the clustering dependency, which misguides the confirmation of the tested hypothesis and the deduction of theories. A substantial body of methodological studies in the early 2000s has highlighted this issue (e.g., Luo & Kowk, 2009; Moerbeek, 2004; Van den Noortgate et al., 2005; Opdenakker & Van Damme, 2000; Tranmer & Steel, 2001). In general, the variances from an omitted level are redistributed into the adjacent levels if the intermediate level is omitted. For example, if the lowest or the highest level is omitted, the variances are distributed to the adjacent higher and lower levels respectively. As the omitted dependencies redistribute to the wrong

cluster levels, the variance estimates of random effects (i.e., variance components) and the standard error estimates of fixed effects (i.e., regression coefficients) are still biased even if the estimated model captures some clustering effects. Also, critical dependencies within the modeled clusters must be represented and accounted for in the model. In other words, all crucial dependencies of clustering should not be falsely left-out or over-specified. The debate of this condition has been mostly around the misspecifications of the error variance-covariance structure of repeated measures in longitudinal data analysis (e.g., in Baek & Ferron, 2013; Ferron et al., 2010; LeBeau, 2018; Murphy & Pituch, 2009).

Ideally, empirical researchers are encouraged to provide the strongest models that are best fit for their data, theories, and research design. On one hand, if any cluster levels are necessary, the corresponding clustering dependencies should be modeled for robust inferences. On the other hand, we do not want to model the unnecessary clustering and fall into the opposite extreme of overcorrection (Abadie et al., 2017; MacKinnon & Webb, 2019).

The first practice is often considered by several conventional criteria for clustering specification (Opdenakker & Van Damme, 2000). A basic one inheres in the conceptualization. To address the substantive research interests relevant to different levels' mechanisms, researchers usually split models into the corresponding multiple levels (Cheong et al., 2001). However, this criterion alone often fails if a cluster level of the mechanism is historically overlooked (Martínez, 2012). Other criteria of defining cluster levels based on the stages of a sampling design and treatment assigned levels in experiment design have been considered (Abadie et al., 2017, 2020; Hedges & Rhoads, 2011; MacKinnon & Webb, 2020; Opdenakker & Van Damme, 2000, Raudenbush, & Schwartz, 2020). However, a researcher may inadvertently omit a cluster level if she ignores the complex sampling structure (Niehaus et al., 2014; Wang et

al., 2019; Zhu et al., 2012; Skinner & Wakefield, 2017). In some cases, the omission of a cluster level is obliged due to data restrictions. For example, many public-available datasets do not provide linkable identification numbers across cluster levels (e.g., classrooms) due to data ethic concerns (Conaway, et al., 2015). Also, it is not surprising that many published studies do not fully illustrate the sampling designs or provide original data. Readers could have reasonable questions of whether the clustering dependencies are modeled correctly.

Conventionally, researchers may also model a cluster level if the size of the clustering dependency measured by the intraclass correlation coefficient (ICC) is considerable. Earlier research suggests a rule of thumb of larger than 0.05 to include a cluster level in modeling. Noticeably, since there are no statistical tests or definite thresholds of ICCs to make a modeling decision, researchers may judge the ICCs based on evidence from previous research. However, evidence from prior research could have different contexts than the current one, thus leading to an inaccurate judgment of the empirical ICCs. Nonetheless, Musca et al. (2011) found that the Type I error rate is always higher than the conventional 5% when clustering is ignored even with an ICC value is as low as 0.01 across many conditions of group size. Therefore, the ICC may not be sufficient for deciding whether a level should be included in analysis. In modeling with more than one level of clustering, judgments based on multiple ICCs may become even more complicated. Alternatively, power analysis of the experiment designs cannot solve the question of how many levels there are in modeling, either. Designed to determine the sample size needed to achieve the power of the statistical hypothesis tests, a power analysis is conducted with a presumption of levels of clustering in design (Berger & Wong, 2009; Cohen, 1992; van Breukelen & Moerbeek, 2016). If a level of clustering matters but is omitted in the design, a

detected educational mechanism or effective intervention may have adequate power, but for an incorrect inference (see Konstantopoulos, 2008a)

The assumptions associated with the second practice of not modeling unnecessary clustering are also often unsatisfied since the current guidelines on when to account for clustering remain vague. For example, analytical guidance would state that a clustering of interest should be accounted for as the estimates change compared with the models without including the cluster level (e.g., Van den Noortgate et al. 2005; Cameron & Miller, 2015). This kind of statement does not explain the rationale of why the cluster gives rise to clustering or when to cluster. This gap causes two major misconceptions. One is that whenever there is a level of cluster that can be defined, regardless of -inherently or by design, a clustering dependency is possible and thus needs to be modeled. Another misconception is that a cluster level needed to be modeled since adding it would change the standard error estimates. It is often the case that empirical researchers choose the larger standard error estimates accounting for clustering dependency to avoid committing a Type I error, without justifying whether the clustering is true and must be adjusted (Robinson, 2020). To dispel these misconceptions, a theoretical framework of when a cluster level is necessary and thus should be controlled is pivotal. This argument has been highlighted in Abadie et al. (2017). Along with Hedges and Rhoads (2011), these studies clarify that the standard error estimates should be corrected if the clustering is due to multi-stage sampling design and randomized experimental by clusters.

1.3 Research Questions and Goals

Despite strong analytical evidence of the risks of omitting levels of clustering and the urgent need for practical guidance to judge whether the estimated model adequately accounts for

clustering, it is still unclear what misspecification of the random effects of the cluster levels may or may not lead to incorrect results. The above discussion motivates the current study to ask the following questions:

1) When should a cluster level and the corresponding clustering dependency to be explicitly modeled, and when could they be omitted?
2) And, if an essential level is omitted in modeling, whether and how much of the omitted clustering dependency would affect the robustness of inference?

This study investigates these questions in settings of a true three-level clustered data structure, while a cluster level, either at the highest, middle, or the lowest level, is omitted in the estimated two-level models. Applying insights from Abadie et al. (2017) and Hedges and Rhoads (2011), the first research question is answered by building a theoretical framework of when a middle or highest cluster level is produced by sampling and experimental designs, but is omitted in modeling. In the omitted lowest cluster level case, the theoretical argument switches to the serial correlation dependency due to the chronologically ordered nature of repeated measures. Answering the first question aids in clarifying empirical decisions of whether a cluster level should or should not be modeled to avoid either Type I or Type II errors and improve the analytical identification of the consequences of an omitted cluster level.

The second question is answered through analytically quantifying the magnitude of the standard error estimates bias of the slope estimates of predictors at each level. Previous studies of examining the issues of omitting a cluster level commonly use simulations to show empirical evidence of bias in estimates and threats to robust inferences. The simulation approach, assuming a known correct model to compare with the other false ones, has advantages in setting extensive

ranges for parameters and models. Nonetheless, though those simulations reveal valuable general patterns, how the bias is produced mathematically is still in a black box. This study complements those simulation-based evidence through closed forms of standard error correction formulas. These formulas, showing the relationship between the bias and the clustering parameters (i.e., ICCs and cluster sample size) of the omitted cluster level, can identify where the omitted clustering is hidden or distributed to other levels and how statistical inferences are affected. In other words, aligning with the theoretical framework already established, the sources of clustering dependency are also clarified. The explanation of the approach is soon introduced in the following Section 1.4.

Finally, with the development of such formulas for standard errors and bias as a function of clustering, this study is further able to develop a sensitivity analysis framework for researchers to quantify the robustness of inferences (or effect size) and the risk of making a false hypothesis decision based on the clustering degree of the suspected omitted cluster level. This sensitivity analysis framework contributes to filling the gaps in current methodological research and to bridging with empirical studies that require guidance in making decisions on modeling specific cluster levels. Particularly, this sensitivity analysis framework is desired in practice when the omitted cluster level is not able to be included in the modeling.

1.4 Combining the Benefits of the Model-Based and the Design-Based Approaches

While the model-based approach HLM explicitly models the multilevel clustering dependency with random effects, the design-based approach provides statistical corrections to the standard error estimates (Cameron & Miller, 2015; Cheong et al., 2001; McNeish & Wentzel, 2017). The design-based approach is prevalent in the fields of survey studies and economics,

where the corrections are called Design Effect (DEFF) and Cluster Robust Standard Errors (CRSE). In a two-level sampling data structure, DEFF is derived from the ratio of the variance of an estimate that takes into account the clustering and the variance that ignores the clustering (Kish, 1995; Snijders, 2005), which is $DEFF = 1 + \rho_{icc} * (\bar{N}_K - 1)$. ρ_{icc} is the ICC and $\bar{N}_K$ is the average cluster size of clusters *k*.

In the field of economics, CRSE is widely applied to many structures of clustering (see a detailed survey in Cameron & Miller, 2015). Generally, CRSE provides a mathematical expression of the variance-covariance structure with an index measuring the error variance (Snijders & Bokser, 2011). In a simplified approximation CRSE case when the homoscedasticity assumption holds[1] (as set in the current study), this index is derived as Moulton Factor (MF) (Angrist & Pischke, 2009; Moulton, 1986, 1990). Moulton Factor is essentially close to DEFF since it is also derived from the ratio of the variance of an estimate with the clustering effect and the variance without the clustering effect[2]. The standard error estimates are corrected by the square root of DEFF or MF**,** which are equivalent to the model-based two-level HLM (Cheong et

[1] The correlated-within-cluster error terms require a covariance matrix estimator that is robust to arbitrary patterns of both heteroskedasticity and intra-cluster correlation (MacKinnon & Webb, 2020). The current work dealing with omitted clustering focuses on the omission of the latter one and assumes homoskedasticity. The setting of homoscedasticity implies that any heteroskedastic patterns in the specified and modeled clustering have been already corrected, and the assumption still hold after included the omitted cluster level. In later chapters of model settings, the cluster sizes of the omitted cluster level are set to be relatively equal, and there are no heterogeneous across clusters. A discussion of modeling heterogeneous random effect variance within an empirical education setting can be seen in Leckie, French, Charlton, and Browne (2014).

[2] Moulton factor is $MF = 1 + \rho_{z,k} * \rho_{icc} * \left(\frac{Var(\bar{N}_K)}{\bar{N}_K} + \bar{N}_K - 1\right)$. Moulton factor uses $\frac{Var(\bar{N}_K)}{\bar{N}_K}$ (i.e., average variance of the cluster size deviation) to account into the variation of unequal cluster sizes. This is equivalent to the Skinner (1986)'s development of Kish's *DEFF*. Additionally, compared with the DEFF, Moulton factor also has $\rho_{z,k}$, which is the within-cluster correlation of the predictor *Z* . When the predictor *Z* is at the aggregated level, this $\rho_{z,k}$ is perfectly correlated and equals to 1, and , thus, MF approaches to *DEFF*. Abadie et al. (2017) argues that $\rho_{z,k} * \rho_{icc}$ may not be sufficient to decide the adjustment, but the "within cluster correlation of the product of the residuals and the regressor" (p.5) (i.e., $\rho_{z,k*error}$) is. In the current study, the uncertainty due to $\rho_{z,k}$ is less of a research interest as $\rho_{z,k} = 1$ for cluster-level predictors.

al., 2001; Huang, 2018; Niehaus et al., 2014). An empirical example showing this equivalency can be seen in Claessens (2012).

The current study considers a three-level clustered data where two layers of clustering are observed while one clustering is omitted in a two-level model, and innovatively applies the method of DEFF to correct for the standard error bias due to the omitted layer of clustering dependency in the estimated two-level HLM model. In this way, closed forms of formulas of quantifying the bias of the standard error estimates can be derived the same as the DEFF. The only difference from the DEFF is that the denominator of the formulas here are the variance estimates from the two-level models, where partial clustering dependency were captured albeit not fully. Cheong et al. (2001) have shown the potential of this idea. Employing a national representative survey data, they compared the standard error estimates from a three-level model with the ones from a two-level model omitting the middle cluster level while having been corrected by CRSE for the two-level HLM estimated model. Those standard error estimates of the later approach that combined model- and design-based approaches are found to be comparable to the empirical standard errors and the ones from the true three-level model.

Current literature has provided other developed approaches to deal with the same issue of omitting a cluster level. For example, Raykov et al. (2016) addresses the question of omitting a middle cluster level through considering the potential size of the middle cluster level variance in the estimation of the confidence intervals of testing the cluster level variances. In Hedges and Rhodes (2011), corrections were made to *F*-test statistics in two-level data while the clustering is omitted. Comparing with these studies, the approach developed in the current study is beneficial in ways of expressing the different sources of clustering dependency as functions of the clustering parameters of random effects variance and sample size of clusters. Further, plugging

in plausible values of the clustering parameters, empirical research can use the sensitivity analysis to evaluate the estimated model and transparently show their analytic decisions (Abe & Gee, 2014)

1.5 Summary of Findings

This study presents closed forms of formulas that quantify the standard error estimation bias due to omitted clustering dependency. A general pattern found is that, if a cluster-level predictor of interest is falsely disaggregated to the lower levels since it is not explicitly modeled, its standard error estimate of the coefficient estimate is underestimated. Specifically, if the middle cluster level is omitted, the middle-level cluster predictor that is disaggregated at the lowest individual level has an underestimated standard error estimate of its coefficient. If the highest cluster level is omitted, the standard error estimate of the coefficient of the highest-cluster level predictor (which is falsely disaggregated at the middle level) is underestimated. Similar patterns apply when the single level OLS are the estimated models.

If the upper adjacent cluster level is omitted, the standard error estimates are overestimated. This pattern is found in the cases where the highest cluster level is omitted, and the standard error estimate of the coefficient defined at the middle cluster level is upward biased, leading to Type II error. In the same vein, though the lowest cluster level predictor is not of the current study's interest, findings show that the corresponding standard error estimates are overestimated if the adjacent higher cluster level is omitted.

An exceptional pattern is that, if the middle cluster level is omitted in the estimated two-level model, the standard error estimate of the highest-level predictor's coefficient is not biased. This is because the overall dependency is captured in the estimated two-level model though the

sources of clustering are entangled. At last, if the omitted level is not adjacent to the level of the predictor of interest, such as when the lowest level predictor is the predictor of interest and the highest cluster level is omitted, the corresponding standard error estimate remains unbiased.

The magnitude of the standard error estimates bias can be calculated by the derived formulas. Furthermore, combined with empirical studies as examples, this study encourages empirical researchers to utilize the developed sensitivity analysis framework to diagnose whether the hypothesized omitted clustering would result in considerable estimation bias that would invalidate any inferences. The sensitivity analysis is of the best usage when the researchers or readers suspect a potential issue of omitting cluster level due to design while there are data restrictions or other reasons that using the model-based approach of modeling that level is not plausible.

1.6 Structure of this Study

This study follows with four chapters, where three chapters (i.e., Chapters 2, 4, and 5) discuss the scenarios of cluster omission respectively at the middle, highest, and lowest level, and one chapter (i.e., Chapter 3) develops the sensitivity analysis framework. In Chapter2, the discussion of omitting the middle cluster levels in the two-level HLM models are based on a theoretical framework of omitted cluster levels due to sampling and experimental designs. For better implication significance, the current study takes the prevalently used national presentative survey datasets initiated by the National Center for Education Statistics (NCES) as examples. In Chapter 3, the sensitivity analysis framework provides three measures of quantifying the inference robustness. An empirical example is provided to demonstrate the sensitivity analysis in testing the inference robustness when a middle cluster level is omitted. The structure of Chapter

4 of discussing the omission of the highest cluster level in the two-level models is identical to that of Chapter 2, including the theoretical framework and the variance inflation factor derivation process, though the specific scenarios and examples of omitting the highest cluster level in sampling and experimental designs are given. Also, an empirical study using the sensitivity analysis framework is presented. Finally, Chapter 5 discusses the case of omitting the lowest cluster level, where the error variance-covariance is misspecified (i.e., omitting the serial correlation in the repeated measures) in the two-level growth modeling.

CHAPTER 2

OMITTED THE MIDDLE CLUSTER LEVEL

2.1 Introduction

The intermediate level reflects important social activities. For example, in educational research, classrooms and teachers, lying between students and schools, contain rich educational activities that largely influence students' daily educational experience within schools (Martínez, 2012; Raudenbush, 2008; Raudenbush & Sadoff, 2008). Empirical studies often employ three-level HLM models to fully reveal the relationships among predictors at the students, classrooms, and school levels (such as Bryk & Raudenbush, 1989; H. C. Hill et al., 2005; Nye et al., 2004). Empirical studies may also choose not to model the middle classroom level, theoretically and methodologically, and conduct two-level models. For example, Martínez (2012) argued that the research field's historical foci on school-level effects might overlook the within-school dynamics of classrooms. In the estimated two-level model, the omitted between-classroom variation is repartitioned into the school- and student-level. The repartition of random effects largely impacts the conclusions drawn for schools since the classrooms often explain a significant proportion of variances that are also far more than what the schools can explain (Martínez, 2012; Beaton & O'Dwyer, 2002). A similar discussion extends to other fields of studies with multilevel social structures as well. For example, Vaezghasemi et al. (2016) found that households, which are between individual and residential communities, are rarely considered when examining the contextual effects on individuals' body mass index.

Still, current literature lacks a practical guide to inform under what circumstances the middle cluster level is necessary to be modeled to represent a complete and accurate educational and social mechanism. This chapter intends to fill this gap by investigating scenarios of omitting a middle cluster level in two-level HLM analyses due to research design. In Section 2.1 and 2.2, those scenarios are classified into mechanisms of two- or three-stage sampling designs and CRT. This classification helps to clarify *when* the middle clustering dependency matters in modeling. Furthermore, this chapter aims to answer how the estimates of predictors at each level would be impacted if the middle cluster level is essential due to design but omitted in modeling. Previous research mainly analyzed the impacts on random effects in unconditional models; this chapter extends the settings to conventional empirical models with predictors and covariates. Section 2.3 details the settings of two- and three-level HLM models with predictors of interest at each level based on the two mechanisms discussed.

To answer the question of *how much* the omitted cluster level matters, this chapter derives mathematical formulas to quantify the estimation bias of random effects and standard errors. The developed formulas adjust the standard error estimates of coefficients defined as variance correction due to omitted clustering (VOC). In a similar format of design effect, VOC is a function of the intraclass correlation coefficient and sample size of the omitted cluster level. It further informs the later sensitivity analysis framework with implications for empirical examples in Chapter 3. A simulation study in Section 2.4 is designed to examine the performance of the bias quantification formulas. Empirically meaningful VOC parameters are selected for the simulation study. Finally, Section 2.5 gives a conclusion.

2.2 Omitted Middle Level Due to Sampling Design

Table 2.1 summarizes when the middle cluster level is omitted in two-level HLM due to sampling design. One scenario is when the SSUs as the middle cluster level is excluded from modeling in a three-stage sampling design; the other scenario is when the omitted middle cluster level appears as incidental instead of being deliberated in a two-stage sampling design. The following considers these two scenarios in a typical educational setting where students are nested within classrooms and classrooms are nested within schools, and a treatment is randomly assigned to schools. The omitted middle level is hypothesized as classrooms and teachers. The current study examines empirical findings that aim to generalize to a broader population of the sampled schools and classrooms, rather than those fixed for the sampled schools and classrooms. In this case, the clustering effects of schools and classrooms matter and need to be considered in modeling (Schochet, 2008; Abadie et al., 2017).

Table 2.1 k-stage sampling design with $k-1$or $k+1$ estimated models.

Sampling Design / **Estimated Model**	**Three-stage Sampling** (e.g., Students – Classroom – Schools)	**Two-stage Sampling** (e.g., Students – Schools)
Two-level Model	Omits the middle classroom cluster level.	Corresponds with the sampling design, while omitting the incidental cluster level.
Three-level Model	Corresponds with the sampling design.	Counts into the incidental cluster levels.

2.2.1 Omitting SSUs in a Three-stage Sampling Structure Data

Consider a dataset that has a three-stage sampling design where schools are PSUs, classrooms are SSUs, and students are USUs. The three-stage design effect accounts for these

two sources of clustering to adjust the standard error estimates (Chen & Rust, 2017; Skinner et al., 1989; Valliant et al., 2013)[3]

$$DEFF_{3L} = 1 + \left(n_{(s2)} - 1\right)\rho_{(s2)} + n_{(s2)}(n_{(s1)} - 1)\rho_{(s1)},$$

where $\rho_{(s1)}$, $\rho_{(s2)}$, $n_{(s1)}$, and $n_{(s2)}$ are correspondingly the first-stage and second-stage intraclass correlation coefficients and average cluster size. Equivalently, a model-based approach, i.e., a three-level HLM model, explicitly analyzes the clustering dependency of students within classrooms and clustering dependency of classrooms within schools.

In practice, the second stage of sampling may be purposively omitted for model simplicity, especially when the substantive research question is not directly related to the middle level (Stapleton & Kang, 2018; Konstantopoulos, 2008a). In the case of omitting the middle cluster level, adapting from the original two-stage sampling design effect in Kish (1995), the design effect of a simplified structure with PSUs of schools and USUs of students disregarding the SSUs of classrooms is $DEFF^*_{2L} = 1 + \left(n^*_{(s1)} - 1\right)\rho^*_{(s1)}$.

The superscript $*$ notes for the setting of SSUs omission. $n^*_{(s1)}$ is the average number of students within a school and equals to the product of $n_{(s1)}$ and $n_{(s2)}$. $\rho^*_{(s1)}$ measures the similarity of students within schools. Figure 2.1 visualizes the intraclass correlation structure of the three-stage sampled data in the upper panel and the one with the omitted SSUs in the lower panel. In the complete structure of a three-stage sampling, $\rho_{(s1)}$ and $\rho_{(s2)}$ capture the within-

[3] The current study assumes no stratification in the sampling design for simplification purposes. However, stratification is commonly used in educational sampling design. For example, schools, as PSUs, are firstly stratified by census units. If the stratification is ignored, Type II error occurs, but is less of a concern when studies prefer conservative results. In Chen and Rust (2017), design effects formulas incorporate stratification with multiple stages.

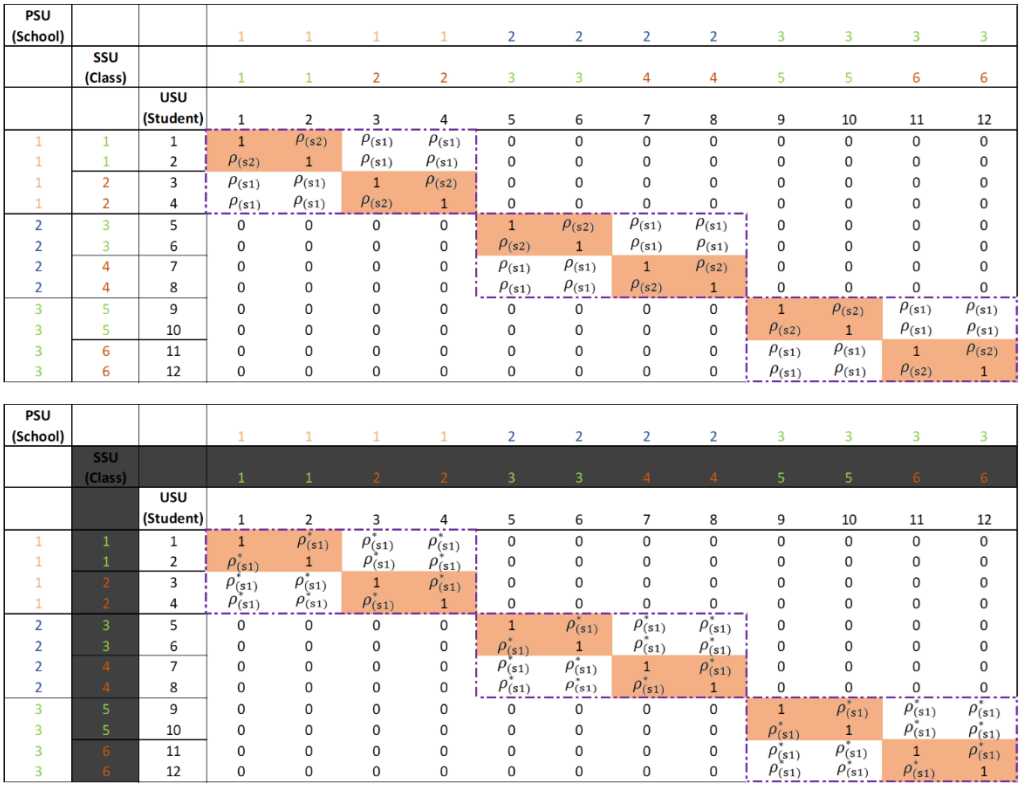


Figure 2.1 Data correlation structures of three-stage sampling designs when the secondary sampling stage is included and omitted

classroom-within-school and between-classroom-within-school clustering dependency within a school PSU as presented in a dashed box. As schools are the randomly sampled PSUs, correlations across schools are zero. When the SSUs are omitted in modeling, as shadowed in the lower panel figure, the within-school dependency is captured by $\rho^*_{(s1)}$, regardless of classrooms. Between-school independency assumption still holds. Intuitively, since the overall clustering dependency remains the same as $DEFF_{3L} = DEFF^*_{2L}$, $\rho^*_{(s1)}$ is a function of $\rho_{(s1)}$, $\rho_{(s2)}$, $n_{(s1)}$ and $n_{(s2)}$.

The existing design effect literature has not been extended to define the mathematical relationship between $DEFF_{3L}$ and $DEFF^*_{2L}$, and the consequences on parameter estimates precision are less known. This unknown relationship can be solved through the model-based

HLM approach through quantifying the relationship of the variance-covariance or intraclass correlation structure of the three-level HLM model and the one of the two-level model. Section 2.3 will detail the solution.

2.2.2 Incidental Middle Level between PSUs and SSUs (or USUs)

Educational datasets commonly provide additional survey data beyond the designed sampling units. For example, NCES datasets, including Early Childhood Longitudinal Study (ECLS), National Assessment of Educational Progress (NAEP), and Education Longitudinal Study (ELS), collected classroom- and teacher surveys to facilitate research to understand the within-school dynamics, even though the sampling designs did not present a known probability sample from classrooms. Wang et al. (2019) defined this scenario as emerging *incidental* middle cluster level in sampling. The cluster levels corresponding to sampling designs, such as the PSUs of schools and SSUs of students, are called *deliberate* levels (McNeish & Wentzel, 2017).

When the sampling design is two-staged structure, such as in NAEP, a two-level model is analytically sufficient to take into account the clustering dependency of students nesting within schools and provides unbiased standard error estimates of the school-level predictors (Cheong et al., 2001; Moerbeek, 2004; Wang et al., 2019). However, the two-level model does not explicitly model the between-classroom variance and does not satisfy research interests that focus on between-classroom variance. Further, the two-level model could completely disregard any potential classroom-level effects or falsely disaggregate the classroom-level predictors at the lower student level. This case may be comparable to the well-documented issue of omitting a single level clustering dependency in a single-level model of OLS estimation. In the setting of disaggregated classroom-level predictors, an artificial homogeneity is introduced at the student level, which produces overestimated standard error estimates of the student-level predictors and

underestimated the standard errors for the classroom-level disaggregated predictors (Korendijk, Hox, et al., 2011). Wang et al. (2019) showed simulation evidence that the standard error estimates of the student-level predictors are unbiased. Section 2.4.3 of the current study further shows that the inconsistency evidence in Wang et al. (2019) is not valid, but due to their parameter setting restrictions.

When the research interests include between-cluster variations at different levels, conducting a three-level model is beneficial since it simultaneously incorporates the sampling stages and the incidental middle cluster level mechanisms. Even in an extreme situation where the between-classroom variations are nearly zero, the estimated variance of the student- and school-level random effects from the three-level model would not be biased (Raykov et al., 2016), though the sampling variance estimates would change slightly due to the changes of degrees of freedom by the added cluster sample size and predictors of classrooms. Nevertheless, Wang et al. (2019) and McNeish et al. (2017) summarized the pitfalls of conducting a three-level model, which are mainly (1) increased complexity of modeling assumptions and the increased risks of violating the assumptions, and (2) the sparseness of the cluster number of the incidental middle level may lead to biased estimates of the variance components. With these concerns and when the research interests only focus on the school-level predictors, a two-level model is preferred[4].

[4] Many studies have provided several solutions to address the second concern of small cluster numbers. McNeish and Stapleton (2014) provided a review of such methods, including restricted maximum likelihood with Kenward–Roger adjustment (see Kenward & Roger, 1997, 2009) and, alternative to maximum likelihood based approaches, Bayesian Markov Chain Monte Carlo (MCMC) (see Baldwin, & Fellingham ,2013; Hox, van de Schoot, & Matthijsse, 2012). However, these discussions mainly focused on addressing the issue within a two-level cluster data structure setting. More studies are needed to extend the discussion in a three-level clustering data structure and examine the methods when the middle cluster level sample size is small. In the current discussion of whether to include an incidental middle cluster level in modeling, the above-mentioned limitations could still affect empirical researchers' modeling decision-making.

Wang et al.(2019) provided modeling guidelines depending on the parameter of interest and listed a few empirical examples which employed the same ECLS data while made different modeling decisions of the incidental cluster level (p. 575). For instance, Jennings and DiPrete (2010) explicitly modeled the incidental teacher-level cluster since their research goal is to examine teacher effects on students' social and behavioral skills. While in Adelson, McCoach, and Gavin (2012), which studied school-level gifted programs' population average effect on student's achievement, the incidental classroom level is not modeled. Their modeling approach is legit since it corresponds to the sampling design that the classroom level is not a sampling stage, and the classroom-level effect is not the focus of the study. Their study also avoided the pitfall of overcorrection if model any unnecessary clustering.

Yet, practical guidelines of modeling choice with incidental middle cluster level have not been widely explored except in Cheong et al. (2001) and Wang et al.(2019). This led to conflicting modeling decisions in empirical research using the same data for similar research questions. For example, Fitchett and Heafner (2017) examined the teacher's professional characteristics and classroom instructions on students' history achievement using the NAEP data. Therefore, even though teachers are not a deliberate sampling stage in NAEP, the authors explicitly modeled the teacher cluster level. However, Heafner, VanFossen, and Fitchett (2019), which employed NAEP as well, conducted a two-level model to examine student characteristics, courses and instructional variables, as well as demographic variables' effects on students' economics content knowledge. The incidental classroom level is suspected to be omitted, and a key predictor of classroom instruction could be a classroom-level variable but falsely aggregated at the student level. Though the school-level predictors' standard error estimates are not biased, the standard error estimates of the key predictors of classroom-level instructions could be

underestimated and the ones of the student-level characteristics could be overestimated. That study will be soon introduced in Chapter 3 to demonstrate the sensitivity analysis.

2.3 Omitted Middle Level in Cluster Randomized Trial

Many CRT design a three-level structure with cluster levels of students, classrooms and teachers, as well as schools, where the randomization happens at the schools and the outcomes are at the student level (Spybrook, Kelcey, et al., 2016; Westine et al., 2013). Two sources of clustering exist in CRT (Schochet, 2008; Abadie et al., 2017): one is the random assignment of units to the control and treatment groups, and the other is the sampling of two-level of clusters from a broader population as discussed in 2.1. In many cases, two-level models with students and schools are conducted, where the clustering due to assignment is captured while clustering due to sampling could be only partially captured. The omission of modeling the classroom level clustering effect could be the result of the two scenarios from sampling design that are discussed above.

Consistent with the previous review, the point estimates of the school-level intervention effect and standard error, as well as the minimum detectable effect size, which are of the most research interest in CRT, are nearly identical in three- and two-level models, regardless of the size of the teacher-level variance, size of clusters, and number of student- and school-level covariates, as evidenced in Murray et al. (1996) and Zhu et al. (2012). Equivalently, the corresponding design effect for the treatment group of schools is the same as the above $DEFF_2^*$, where the overall clustering dependency within schools is captured (Hedges & Rhoads, 2011). However, the potential classroom-level effects and cross-level moderation effects of the intervention are ignored, which are pivot in CRT studies that aims to detect heterogeneous

treatment effects and answer questions of *how* and *under what conditions* the intervention works beyond *what* works (Spybrook et al., 2016; Spybrook, Zhang, et al., 2020).

Recently, scholars call for advancing the understanding of the implementation process of interventions in school settings, such as how teachers deliver the treatment to students (Lendrum & Humphrey, 2012). For example, teachers may be influenced by the local contexts and adapt the intervention process, and students could be assigned to teachers based on certain attributes of teachers, such as the experience of teaching or class schedule(Weiss, 2010; Weiss et al., 2016). Also, it is not uncommon that teachers are often trained as groups for the intervention (such as in Jayanthi et al., 2018) that groups of teachers may conduct the intervention similarly. These situations result in students who have the same teacher or are exposed to a teacher group could receive the treatment in a similar manner. Therefore, if the CRT design considers the role of teachers, the correlation of treatment and clustering in CRT would be a composition of treatment correlating within teachers (or teacher groups) and schools.

In the work of Abadie et al. (2017), potential treatment provider variation is mentioned while considering the classroom and teacher level effects as fixed rather than intending to generalize the effects to the superpopulation of classrooms and teachers. The current study, on the contrary, considers the classrooms as SSUs in a three-stage sampling design or as an incidental cluster level that is not in a sampling stage. The current work also explores the influence on coefficients associated with students and teacher level predictors when the between-teacher variance is omitted in a two-level model, which is not studied in Zhu et al. (2012).

2.4 Quantification of Standard Error Bias

This section formulates the potential bias of the standard error estimates of predictors when a middle cluster level is omitted. The process of quantifying the bias is, in essence, a design-based approach, which compares the variance estimates from a satisfactory three-level random intercept model and an estimated two-level random intercept model. The models are set to cover the previously discussed scenarios of omitting the middle clusters due to sampling and experimental designs. Meanwhile, the notation used throughout the whole study is explained.

2.4.1 Model Setting

Two-level random intercept model

Consider first a two-level random intercept model with a continuous dependent variable Y_{ik}, which indicates the outcome of a student *i* in a school *k*. The model is:

$$\text{Student-level: } Y_{ik} = \beta_{0k} + \beta_{1k}X_{ik} + \beta_{2k}W_{ik} + \tilde{\varepsilon}_{ik},$$

$$\text{School-level: } \beta_{0k} = \gamma_{00} + \gamma_{01}\, Z_k + \tilde{u}_{0k},$$

$$\beta_{1k} = \gamma_{10},$$

$$\beta_{2k} = \gamma_{20},$$

$$\text{Mixed model: } Y_{ik} = \gamma_{00} + \gamma_{10}X_{ik} + \gamma_{20}W_{ik} + \gamma_{01}\, Z_k + \tilde{u}_{0k} + \tilde{\varepsilon}_{ik}.$$

X_{ik} and W_{ik} are treated as student-level predictors. X_{ik}, for instance, can be the prior scores of students, which is a commonly used student-level covariate (e.g., Bloom et al., 2007). However, W_{ik} is actually a classroom-level measure, such as an attribute of teacher, so that all students in the same class have the same value of W_{ik}. This setting is to satisfy the falsely

disaggregated incidental cluster level predictor case. The random intercept β_{0k} is predicted by a school-level predictor Z_k to capture the variability between schools. The setting of Z_k being either a continuous variable or a binary treatment predictor as in CRT does not affect the later quantification process of the potential bias of variance estimates. The latter section soon confirms this note. Additionally, the predictors are assumed to be orthogonal to the random effects at any level for the exogeneity assumption because X_{ik} and W_{ik} are group-mean centered (Antonakis et al., 2019). To keep the simplicity of the conceptual example, I only present each level with the minimum number of predictors, albeit many other covariates can be added. As long as the assumptions hold, the following algebraic expressions of the variance estimation and the quantification procedure of bias remain the same.

Conventionally, the random effects are assumed to be normally distributed with means of zero and constant variances conditioning on the predictors and have zero covariance, which are$\tilde{\varepsilon}_{ik} \sim N\left(0, \tilde{\sigma}^{(i)}\right)$, $\tilde{u}_{0k} \sim N\left(0, \tilde{\sigma}^{(k)}\right)$, and $cov(\tilde{\varepsilon}_{ik}, \tilde{u}_{0k}) = 0$. Tildes over the parameters are used to distinguish the current two-level model from the later three-level model. The total sample size of students is $M_K * n_0$, where M_K is the number of schools, and n_0 is the average number of students within schools[5].

For each school *k*, the error variance-covariance matrix of Y_k, denoted as $\widetilde{\boldsymbol{\psi}}_K$, is composed of a residual variance matrix $\widetilde{\boldsymbol{R}}$ and a random intercept variance matrix $\widetilde{\boldsymbol{G}}$:

$$\widetilde{\boldsymbol{\psi}}_K = var(Y_k) = \widetilde{\boldsymbol{R}} + \boldsymbol{l}_{n_0} \widetilde{\boldsymbol{G}} \boldsymbol{l}'_{n_0},$$

[5] The current study considers balanced design as a starting point, where the cluster size is assumed to be the same (or almost identical) across cluster units. This setting provides closed forms of maximum likelihood estimates . Thus, I can make comparisons of the estimates across two- and three-level models and the OLS models in later chapters, when fixing the coefficient estimates of the HLM and OLS to be equal (Nezlek & Zyzniewski, 1998). Van den Noortgate et al. (2005) has provided simulation evidence of omitting a cluster level in unbalanced settings. They found similar patterns of the variance-covariance repartition as in balanced settings.

where the dimension of $\widetilde{\boldsymbol{\psi}}_K$ is n_0 by n_0, and $\boldsymbol{l}_{n_0}$ is a column vector of n_0 ones.

Further,

$$\widetilde{\boldsymbol{R}} = \tilde{\sigma}^{(i)}\boldsymbol{I} = \begin{bmatrix} \tilde{\sigma}^{(i)} & 0 & \cdots & 0 \\ 0 & \tilde{\sigma}^{(i)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\sigma}^{(i)} \end{bmatrix}$$

and

$$\boldsymbol{l}_{n_0}\widetilde{\boldsymbol{G}}\boldsymbol{l}_{n_0}{}' = \begin{bmatrix} \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(k)} \\ \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(k)} \end{bmatrix}.$$

Also, $\widetilde{\boldsymbol{\psi}}_K$ can be write as:

$$\widetilde{\boldsymbol{\psi}}_K = \widetilde{\boldsymbol{R}} + \boldsymbol{l}_{n_0}\widetilde{\boldsymbol{G}}\boldsymbol{l}'_{n_0} = \begin{bmatrix} \tilde{\sigma}^{(i)} + \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(k)} \\ \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(i)} + \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\sigma}^{(k)} & \tilde{\sigma}^{(k)} & \cdots & \tilde{\sigma}^{(i)} + \tilde{\sigma}^{(k)} \end{bmatrix}$$

$$= \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

$$= \sigma^2\left[(1-\rho)\boldsymbol{I} + \rho\, \boldsymbol{l}_{n_0}\boldsymbol{l}'_{n_0}\right], \qquad (2.1)$$

where $\sigma^2 = \tilde{\sigma}^{(i)} + \tilde{\sigma}^{(k)}$ is the total error variance, and $\rho = \frac{\tilde{\sigma}^{(K)}}{\sigma^2} = corr(y_{ik}, y_{i'k})$ is the proportion of variance at the school level[6] or the intraclass correlation coefficient indicating the expected correlation of any two randomly drawn students in a school. The structure of $\tilde{\boldsymbol{\psi}}_K$ is consistent with the purple dashed boxes in the lower panel of Figure 2.1. With new notations of ICCs, Figure 2.2 below modifies Figure 2.1 to show the correlation structure of $\tilde{\boldsymbol{\psi}}_K$ (as shown in the lower panel) and $\boldsymbol{\psi}_K$ of the three-level model(as shown in the upper panel) in the following discussion.

School K			1	1	1	1	2	2	2	2	3	3	3	3
	Class J		1	1	2	2	3	3	4	4	5	5	6	6
		Student I	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	ρ_1	ρ_2	ρ_2	0	0	0	0	0	0	0	0
1	1	2	ρ_1	1	ρ_2	ρ_2	0	0	0	0	0	0	0	0
1	2	3	ρ_2	ρ_2	1	ρ_1	0	0	0	0	0	0	0	0
1	2	4	ρ_2	ρ_2	ρ_1	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	ρ_1	ρ_2	ρ_2	0	0	0	0
2	3	6	0	0	0	0	ρ_1	1	ρ_2	ρ_2	0	0	0	0
2	4	7	0	0	0	0	ρ_2	ρ_2	1	ρ_1	0	0	0	0
2	4	8	0	0	0	0	ρ_2	ρ_2	ρ_1	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	ρ_1	ρ_2	ρ_2
3	5	10	0	0	0	0	0	0	0	0	ρ_1	1	ρ_2	ρ_2
3	6	11	0	0	0	0	0	0	0	0	ρ_2	ρ_2	1	ρ_1
3	6	12	0	0	0	0	0	0	0	0	ρ_2	ρ_2	ρ_1	1

School K			1	1	1	1	2	2	2	2	3	3	3	3
	Class J		1	1	2	2	3	3	4	4	5	5	6	6
		Student I	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	ρ	ρ	ρ	0	0	0	0	0	0	0	0
1	1	2	ρ	1	ρ	ρ	0	0	0	0	0	0	0	0
1	2	3	ρ	ρ	1	ρ	0	0	0	0	0	0	0	0
1	2	4	ρ	ρ	ρ	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	ρ	ρ	ρ	0	0	0	0
2	3	6	0	0	0	0	ρ	1	ρ	ρ	0	0	0	0
2	4	7	0	0	0	0	ρ	ρ	1	ρ	0	0	0	0
2	4	8	0	0	0	0	ρ	ρ	ρ	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	ρ	ρ	ρ
3	5	10	0	0	0	0	0	0	0	0	ρ	1	ρ	ρ
3	6	11	0	0	0	0	0	0	0	0	ρ	ρ	1	ρ
3	6	12	0	0	0	0	0	0	0	0	ρ	ρ	ρ	1

Figure 2.2 Correlation structures of $\boldsymbol{\psi}_K$ of the three-level model and $\tilde{\boldsymbol{\psi}}_K$ of the two-level model omitting the middle cluster level.

[6] The intraclass correlation coefficient can be conditioned on the predictors. For simpler notation, I do not put additional subscript (such as $\tilde{\sigma}^{(i)}_{adj.}$ or $\rho_{adj.}$) to indicate this essence.

Three-level Random Intercept Model

If there emerges a necessary classroom-level middle cluster, a three-level model for students *i* within classroom *j* within school *k* should be conducted as:

$$\text{Student-level: } Y_{ijk} = \pi_{0jk} + \pi_{1jk}X_{ijk} + \varepsilon_{ijk},$$

$$\text{Classroom-level: } \pi_{0jk} = \beta_{00k} + \beta_{01k}W_{jk} + \gamma_{0jk},$$

$$\pi_{1jk} = \beta_{10k},$$

$$\text{School-level: } \beta_{00k} = \gamma_{000} + \gamma_{001}\,Z_k + u_{00k},$$

$$\beta_{01k} = \gamma_{010},$$

$$\beta_{10k} = \gamma_{100},$$

$$\text{Mixed model: } Y_{ijk} = \gamma_{000} + \gamma_{100}X_{ijk} + \gamma_{010}W_{jk} + \gamma_{001}\,Z_k + u_{00k} + \gamma_{0jk} + \varepsilon_{ijk}.$$

Compared with the above two-level model, the three-level model has an additional classroom-level random effect γ_{0jk} which indicates variability across teachers within schools. The predictor W_{jk} is now correctly specified at the middle cluster level to explain the outcome mean differences across teachers within schools. The random effects are assumed to be normally distributed with means of zero and constant variances, which are $\varepsilon_{ijk} \sim N\left(0, \sigma^{(i)}\right)$, $r_{0jk} \sim N\left(0, \sigma^{(j)}\right)$, and $u_{00k} \sim N\left(0, \sigma^{(k)}\right)$. Also, the random effects have zero covariance with each other.

The sample size of schools M_K and the total sample size of students (i.e., $M_K * n_0$) remain the same, regardless of adding or omitting the middle classroom cluster level. In the

three-level model, n_L is the cluster size of the lower nesting level (i.e., the average class size or the number of students taught by each teacher), and n_H is the cluster size of the higher nesting level (i.e., the average number of teachers within each school). Also, n_0 is the average school size or the average number of students within a school.

The following $\boldsymbol{\psi}_K$ is the error variance-covariance matrix of a school *k*, which has a consistent structure as shown in the purple dashed boxes of the upper panel in Figures 2.1 and 2.2. As shown, $\boldsymbol{\psi}_K$ and $\widetilde{\boldsymbol{\psi}}_K$ have the same dimensionality of n_0 by n_0, while, since the single nesting structure in the two-level model is now extended to two levels of nesting, the dimension of $\boldsymbol{\psi}_K$ in the three-level model becomes $(n_L * n_H) * (n_L * n_H)$ as $n_0 = n_L * n_H$. And,

$$\boldsymbol{\psi}_K = \sigma^2 \begin{bmatrix} \boldsymbol{\omega}_J & \boldsymbol{\rho_2} & \cdots & \boldsymbol{\rho_2} \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{\rho_2} & \boldsymbol{\rho_2} & \boldsymbol{\omega}_J & \boldsymbol{\rho_2} \\ \boldsymbol{\rho_2} & \boldsymbol{\rho_2} & \cdots & \boldsymbol{\omega}_J \end{bmatrix}.$$

Its diagonal element $\boldsymbol{\omega}_J$ is $n_L * n_L$in dimension and is the highlighted area within each purple dashed box in the upper panel of Figure 2.2.

$$\boldsymbol{\omega}_J = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_1 \\ \rho_1 & 1 & \cdots & \rho_1 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_1 & \rho_1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} 1 & \rho_2 + \rho_0 & \cdots & \rho_2 + \rho_0 \\ \rho_2 + \rho_0 & 1 & \cdots & \rho_2 + \rho_0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_2 + \rho_0 & \rho_2 + \rho_0 & \cdots & 1 \end{bmatrix}.$$

The off-diagonal element ρ_1 in $\boldsymbol{\omega}_J$ is the intraclass correlation coefficient of any two students from the same classroom *j* in a school *k*, and $\rho_1 = \frac{\sigma^{(k)} + \sigma^{(j)}}{\sigma^2} = \frac{\sigma^{(k)}}{\sigma^2} + \frac{\sigma^{(j)}}{\sigma^2} = \rho_2 + \rho_0$. Intuitively, ρ_1 combines the similarity of students exposed by being in the same school *k* and the similarity of students exposed by being in the same classroom *j*. Specifically, within the school *k* between classrooms, students' similarities are measured by ρ_2. And the average correlation of

any two students from the same classroom is ρ_0. Other ways of defining the intraclass correlation coefficient exists; and Appendix 2.A compares these approaches and presents the derivation of ρ_1.

$\boldsymbol{\rho_2}$ in $\boldsymbol{\psi}_K$ shows the proportion of between-classroom-within-school variation, which is the unhighlighted parts within any purple dashed boxes in the upper panel of Figure 2.2. $\boldsymbol{\rho_2}$ has a dimension of $(n_0 - n_L) * (n_0 - n_L) = n_L(n_H - 1) * n_L(n_H - 1)$, and

$$\boldsymbol{\rho_2} = \begin{bmatrix} \rho_2 & \rho_2 & \cdots & \rho_2 \\ \rho_2 & \rho_2 & \cdots & \rho_2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_2 & \rho_2 & \cdots & \rho_2 \end{bmatrix}.$$

As shown, the expected correlation among students coming from the same classroom (i.e., ρ_1) is larger than the expected correlation among students coming from the same school but different classrooms (i.e., ρ_2); and this difference is measured by ρ_0. In the estimated two-level models, this similarity difference among different cluster levels is ignored. Finally, since the schools as PSUs are the highest cluster level and are independent to each other, the correlations among students from different schools are set to be 0, as shown in the cells outside of all dashed purple boxes in Figure 2.2.

With some algebraic operations, $\boldsymbol{\psi}_K$ can be written as:

$$\boldsymbol{\psi}_K = \sigma^2\{\boldsymbol{I_K} \otimes [(1 - \rho_1)\boldsymbol{I}_J + (\rho_1 - \rho_2)\boldsymbol{l}_J\boldsymbol{l'}_J] + \rho_2\, \boldsymbol{l}_{n_0}\boldsymbol{l'}_{n_0}\}, \tag{2.2}$$

where $\boldsymbol{I_K}$ is an n_H by n_H diagonal matrix, and $\boldsymbol{I_J}$ is an n_L by n_L diagonal matrix. Additionally, $\boldsymbol{l}_J$ and $\boldsymbol{l}_{n_0}$ are vector columns of n_L and n_0 ones, respectively.

Evidenced in Moerbeek (2004), Tranmer and Steel (2001), and Konstantopoulos (2007), the random effects' variances of the three-level model can be approximately repartitioned by the ones of the two-level models. Specifically, the omitted teacher-level variance in the two-level model is partially distributed to the flanking student and school levels as:

$$\tilde{\sigma}^{(i)} \cong \sigma^{(i)} + (1-\eta)\sigma^{(j)}, \tag{2.3}$$

$$\tilde{\sigma}^{(k)} \cong \sigma^{(k)} + \eta\sigma^{(j)}, \tag{2.4}$$

and

$$\eta = \frac{n_L - 1}{n_L * n_H - 1} = \frac{n_L - 1}{n_0 - 1}. \tag{2.5}$$

Thus, the ratio of classroom size to the school size (i.e., η) decides the extent of repartition of the omitted classroom-level variance into the student- and school-level variance. $\eta = \frac{n_L - 1}{n_0 - 1}$ is restricted to $[0, 1]$ since $n_L \leq n_0$ and n_H is an integer that is larger than or equal to 1. When $\eta = 0$, $n_L = 1$ and $n_H = n_0 \neq n_L$, each classroom SSU has only one sampled student, then the between-classroom variance $\sigma^{(j)}$ is dominated by the estimated student-level variance $\tilde{\sigma}^{(i)}$ in the two-level model. When $\eta = 1$, $n_L = n_0$ and $n_H = 1$, all sampled students come from the only classroom SSU in a school PSU, then the between-classroom-within-school $\sigma^{(j)}$ is actually 0 that the estimated two-level model is appropriate.

Figure 2.3 below shows the range of η in an example setting of class size $n_L \in [1, 50]$ and the school size $n_0 \in [100, 500]$. This restriction is due to $\tilde{\sigma}^{(k)} \cong \sigma^{(k)} + \eta\sigma^{(j)}$, where $\sigma^{(j)} > 0$, $\sigma^{(k)} > 0$, and $\eta \geq 0$. In practice, defining the value of η needs to consider this restriction in setting empirical meaningful random effects variance. For example, the value of η decides the

maximum value of $\sigma^{(j)}$ that a researcher can set to satisfy the conditions of $\sigma^{(k)} \in [0, \tilde{\sigma}^{(k)}]$ and $\sigma^{(j)} > 0$, when fixing n_o, n_L, and $\tilde{\sigma}^{(k)}$(or ρ). This discussion will be further shown in the empirical study example of implementing the sensitivity analysis in Chapter 3.

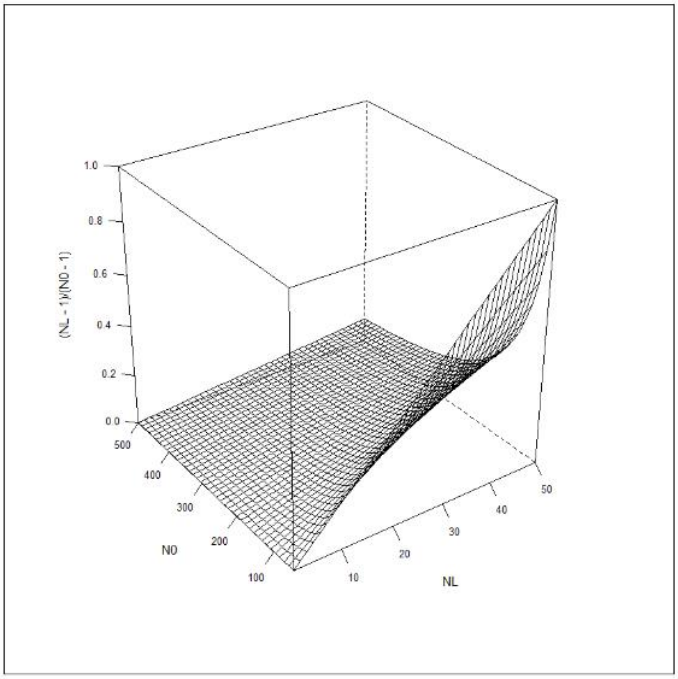


Figure 2.3 Relationship among η, n_0, and n_L

The original ρ in the two-level model now turns into a function of the two intraclass correlation coefficients of the three-level model, which is $\rho = \frac{\tilde{\sigma}^{(k)}}{\sigma^2} = \frac{\sigma^{(k)} + \eta\sigma^{(j)}}{\sigma^2} = \rho_2 + \eta\rho_0$. Further, $\rho_2 = \rho - \eta\rho_0$, and $\rho_1 = \rho_2 + \rho_0 = \rho + (1 - \eta)\rho_0$[7]. Thus, if the classroom-level cluster is omitted, the estimated within-classroom student correlation is upwardly biased by $\rho_1 - \rho = (1 - \eta)\rho_0$, and the between-classroom-within-school student correlation is downwardly biased by $\rho_2 - \rho = -\eta\rho_0$.

[7] The current paper assumes that homogeneity assumption still holds when the teacher cluster level is included in the three-level model. Particularly, when the cluster sizes are equal (or at least has relatively small variances across clusters), the repartitioned variances, though their values depend on the size of η, remain constant across groups.

Throughout the whole study, I use the term *satisfactory model* to refer to the three-level model stated above as it satisfies the specifications of three clustering levels and corresponding random effects. I then name the two-level model omitting a necessary cluster level as the estimated model. Table 2B.1 in Appendix 2B summarizes and compares the model specification, assumption, and estimation considerations and settings of the two-level estimated model omitting the middle cluster level and the three-level satisfactory model. As shown, the only distinctions between the two-level and three-level models occur at the random effect specifications of cluster levels and the allocation of the omitted cluster level's predictor. These distinctions due to omitting cluster levels are the research focus of the current study. Other model assumptions and specifications are conventional settings. Discussions of how to deal with violations of those conventional assumptions in practice are out of the current study's scope. Some techniques to correct for violations of those conventional assumptions (such as small cluster size) are noted in footnotes. In the later section of Discussion, some assumptions (such as balanced design and no random slope) that are closely related to the random effect specifications are considered as limitations and directs for future studies.

2.4.2 Quantifying the Standard Error Estimate Bias

Bias of the Standard Error Estimates of the Coefficients of $\boldsymbol{Z_{(j)k}}$ and $\boldsymbol{Z_{i(jk)}}$

In the two-level model, the estimated variance of the coefficient of $\boldsymbol{Z_{(j)k}}$ is:

$$Var_{2L}(\gamma_{01}) = \left\{\sum_{k=1}^{M_K}(Z_{(j)k}{}'\,\widetilde{\boldsymbol{\psi}}_{\boldsymbol{K}}Z_{(j)k})\right\}^{-1} = \sigma^2\tilde{\tau}_{Z_{(j)k}}\left\{\sum_{k=1}^{M_K}(Z_{(j)k}{}'\,Z_{(j)k})\right\}^{-1},$$

where $\tilde{\tau}_{Z_{(j)k}}$ =$1 + (n_o - 1)\rho$. In the CRT setting, $Var_{2L}(\gamma_{01})$ is the variance of the intervention effect or, in other words, the variance of the group mean difference in outcome such that $Var_{2L}(\gamma_{01}) = Var(\bar{Y}_{ik,1} - \bar{Y}_{ik,0})$. The standard error estimate is the square root of the diagonal of the variance matrix $Var_{2L}(\gamma_{01})$. The subscript $2L$ indicates the two-level model. In a single-level analysis with OLS estimation, the variance estimate is $Var_{OLS}(\gamma_Z) = \sigma^2({Z_{(\mathrm{k})}}'Z_{(\mathrm{k})})^{-1}$.

Compared with $Var_{2L}(\gamma_{01})$, $Var_{OLS}(\gamma_Z)$ is smaller and thus leads to Type I error. The ratio of $Var(\gamma_{01})$ and $Var_{OLS}(\gamma_Z)$ is $\tilde{\tau}_{Z_{(j)k}}$, which is known as the design effect of a two-stage sampling design in survey studies. It quantifies the variance inflation or the over-estimated precision of the effect of $Z_{(j)k}$ as if the sampling scheme is a simple random sample. Or in economics, $\tilde{\tau}_{Z_{(j)k}}$ is the MF that is robust to clustering but assumes homoskedasticity. The detailed derivation procedure of $\tilde{\tau}_{Z_{(j)k}}$ can be found in Angrist and Pischke (2008) and Cameron and Miller (2015).

Similarly, when Z_k is modeled in the three-level model, the variance estimate yields to:

$$Var_{3L}(\gamma_{001}) = \left\{\sum_{k=1}^{M_K}({Z_K}'\boldsymbol{\psi_K}\,Z_K)\right\}^{-1} = \sigma^2\tau_{Z_k}\left\{\sum_{k=1}^{M_K}({Z_k}'\,Z_k)\right\}^{-1},$$

where $\tau_{Z_k} = 1 + (n_L - 1)\rho_1 + n_L(n_H - 1)\rho_2$. Again, in CRT, $Var_{3L}(\gamma_{001}) = Var(\bar{Y}_{ijk,1} - \bar{Y}_{ijk,0})$. The index τ_{Z_k} is derived from the error variance-covariance matrix $\boldsymbol{\psi}_K$ of the three-level model (i.e., Eq. 2.2), which is identical to the three-stage sampling design effect formulas shown in Chen and Rust (2017), and an earlier three-level clinical CRT design work in Heo and Leon (2008). Algebraically, the derivation of the weighting indices of τ_{Z_k} and $\tilde{\tau}_{Z_{(j)k}}$ is straightforward

that τ_{Z_k} and $\tilde{\tau}_{Z_{(j)k}}$ are equal to the summation of all the intraclass correlation coefficients in the brackets of $\boldsymbol{\psi}_K$ and $\widetilde{\boldsymbol{\psi}}_K$, respectively. This procedure implies that $\tau_{Z_k} = \tilde{\tau}_{Z_{(j)k}}$ since all between-cluster variance is captured. However, one must still determine in which cluster levels of the between-cluster variance exists. In essence, τ_{Z_k} takes into account the inflation due to the dependency of two levels of nesting (i.e., students nested within teachers, and teachers nested within schools), compared with $\tilde{\tau}_{Z_{(j)k}}$ which quantifies the variance inflation due to the dependency of a single level nesting (i.e., students nested within schools). The following provides additional algebraic proofs.

The index quantifies the bias of the variance estimate of Z_{k}'s coefficient due to the omitted middle cluster level is the ratio of $Var_{3L}(\gamma_{001})$ and $Var_{2L}(\gamma_{01})$:

$$VOC_M^{(3-2,2L)} = \frac{Var_{3L}(\gamma_{001})}{Var_{2L}(\gamma_{01})} = \frac{\tau_{Z_k}}{\tilde{\tau}_{Z_{(j)k}}}.$$

VOC stands for the Variance bias due to the Omitted Cluster level. The superscript (*3-2, 2L*) indicates that the predictor of interest is at level-3 but modeled as level-2 in a two-level cluster structure. The subscript *M* stands for the omitting of the middle cluster level case. The construction of $VOC_M^{(3-2,2L)}$ follows the same logic of DEFF and MF, which is comparing the variance estimates with and without the omitted cluster level.

In practice, researchers can compute the variance inflation magnitude by filling the possible values of class size n_L , and the average correlation of students from the same class ρ_1. Therefore, I re-express all the variance inflation factors by the known ρ from the estimated two-level model and the assumed omitted level clustering parameters of ρ_1 and n_L. Further, since

$$\tau_{Z_k} = 1 + (n_L - 1)\rho_1 + n_L(n_H - 1)\rho_2$$
$$= 1 + (n_L - 1)[\rho + (1 - \eta)\rho_0] + n_L(n_H - 1)(\rho - \eta\rho_0)$$
$$= 1 + (n_o - 1)\rho = \tilde{\tau}_{Z_{(j)k}},$$

then,

$$VOC_M^{(3-2,2L)} = \frac{\tau_{Z_k}}{\tilde{\tau}_k} = 1. \quad (2.6)$$

$VOC_M^{(3-2,2L)} = 1$ suggests that the estimated variance of the fixed effect of school-level predictor Z_k does not need any bias correction when the teacher-level cluster is omitted. Since the omitted teacher-level variance is redistributed to the school- and student- level, $\tilde{\boldsymbol{\psi}}_{\boldsymbol{K}}$ from the two-level model still takes into account the between-teacher variance.

Equivalently showing in the CRT settings, assuming half schools are randomly assigned to the treatment and control groups (i.e., the sample size of the treatment and control groups is $M_K/2$) , the standard error estimates of γ_{001} and γ_{01} equations in the three- and two-level CRT balanced design (Konstantopoulos, 2008a; Spybrook et al., 2016) are respectively defined as:

$$SE(\bar{Y}_{ijk,1} - \bar{Y}_{ijk,0}) = \sqrt{\frac{4}{M_K n_H n_L}}\sqrt{n_H n_L \sigma^{(k)} + n_L \sigma^{(j)} + \sigma^{(i)}},$$

and

$$SE(\bar{Y}_{ik,1} - \bar{Y}_{ik,0}) = \sqrt{\frac{4}{M_K n_0}}\sqrt{n_0 \tilde{\sigma}^{(k)} + \tilde{\sigma}^{(i)}}.$$

Plugin Eqs. 2.3 – 2.5 of the algebraic relationships between $\sigma^{(k)}$ and $\tilde{\sigma}^{(k)}$, and $\sigma^{(i)}$ between $\tilde{\sigma}^{(i)}$, the standard error estimates of γ_{001} and γ_{01} are equal as shown below:

$$SE\left(\bar{Y}_{ijk,1} - \bar{Y}_{ijk,0}\right) = \sqrt{\frac{4}{M_K n_H n_L}} \sqrt{n_H n_L (\tilde{\sigma}^{(k)} - \eta\sigma^{(j)}) + n_L \sigma^{(j)} + (\tilde{\sigma}^{(i)} - (1-\eta)\sigma^{(j)})}$$

$$= \sqrt{\frac{4}{M_K n_0}} \sqrt{n_0 \tilde{\sigma}^{(k)} + \tilde{\sigma}^{(i)}} = SE(\bar{Y}_{ik,1} - \bar{Y}_{ik,0}).$$

Therefore, if the predictor of interest is at the highest school level, either a binary treatment or a continuous variable, the corresponding fix effect's standard error estimate is unbiased even if the teacher-level variance is omitted, assuming there are no even higher cluster levels than schools. This finding is consistent with Wang, et al. (2019), Zhu et al. (2012), and Cheong et al. (2001).

Extending to an extreme case where the clustering structure is completely ignored as in a single-level analysis with OLS estimation, the variance estimates of $Z_{i(jk)}$'s coefficient needs an adjustment of:

$$VOC_M^{(3-1,OLS)} = \tau_{Z_k} = 1 + (n_L - 1)\rho_1 + n_L(n_H - 1)\rho_2$$

$$= \tilde{\tau}_{Z_{(j)k}} = 1 + (n_o - 1)\rho. \qquad (2.7)$$

Constructed by dividing $Var_{3L}(\gamma_{01})$ by $Var_{OLS}(\gamma_Z) = \sigma^2(Z_{(\mathrm{k})}{}'Z_{(\mathrm{k})})^{-1}$, $VIF_M^{(3-1,OLS)}$ reflects the two-layer nesting structure of $\boldsymbol{\psi_K}$. The magnitude of adjustment depends on the clustering parameters of intraclass correlation coefficients and cluster sizes. Further, $VIF_M^{(3-1)}$ is also equivalent to $\tilde{\tau}_{Z_{(j)k}}$, which captures the total between-cluster variance, while blurring the levels of clustering structure.

Bias of the Standard Error Estimates of the Coefficients of $W_{i(j)k}$ and $W_{i(jk)}$

The following discussion switches to the teacher-level predictor W_{jk} which is falsely aggregated at the lowest level of students as $W_{i(j)k}$, omitting the teacher-level variance $\sigma^{(j)}$ in an estimated two-level model. The inflation of the variance estimate of W_{jk}'s coefficient γ_{010} is quantified similarly as above, though the focus shifts from the two-layer clustering to the single-layer omitted clustering of students nested within teachers. In this simplification, the true error variance-covariance structure only needs to consider $\boldsymbol{\omega}_J$ from $\boldsymbol{\psi}_K$ of the three-level model instead of the whole structure of $\boldsymbol{\psi}_K$. This true variance estimate of W_{jk}'s coefficient γ_{010}, denoted as $\widetilde{Var}_{3L}(\gamma_{010})$, produces a variance weighting index $\tau_{W_{jk}} = 1 + (n_L - 1)\rho_1$.

Further, dividing $\widetilde{Var}_{3L}(\gamma_{010})$ by the variance estimate $Var_{OLS}(\gamma_W) = \sigma^2\boldsymbol{I}$ of the single-level analysis with OLS estimation which falsely assuming students are independent within classrooms and schools, the variance inflation measure yields to

$$VOC_M^{(2-1,2L)} = \frac{\widetilde{Var}_{3L}(\gamma_{010})}{Var_{OLS}(\gamma_W)} = \tau_{W_{jk}} = 1 + (n_L - 1)\rho_1 = 1 + (n_L - 1) * (\rho_2 + \rho_0), \quad (2.8.1)$$

which contains the between-school variance ($\sigma^{(k)}$) and between-classroom variance ($\sigma^{(j)}$) in a correctly specified three-level model. Further, $VOC_M^{(2-1,2L)}$ can be rewritten as a function of the known ρ from the estimated two-level model and the unknown n_L and ρ_0 that researchers can specify as

$$VOC_M^{(2-1,2L)} = 1 + (n_L - 1) * [\rho + (1 - \eta)\rho_0]. \quad (2.8.2)$$

Intuitively, the bracket quantifies the omitted clustering dependency, which consists of (1) the overestimated school-level variance, as presented by $\rho = \frac{\tilde{\sigma}^{(K)}}{\sigma^2}$, from the estimated two-level model, and (2) the uncaptured classroom-level variance $\rho_0 = \frac{\sigma^{(j)}}{\sigma^2}$. Noticeably, these two components are weighted by n_L, and relevantly, η. When $n_L = 1$ that each sampled classroom within a school has only one sampled student, no adjustment is needed for the coefficients' standard error estimates since the classroom-level predictor actually measures the singleton sampled student, which thus can be disaggregated at the student level. When $\rho_0 = 0$, the school-level variance estimate $\tilde{\sigma}^{(K)}$ is not overestimated and equals to $\sigma^{(k)}$. In this case, the estimated two-level model is satisfactory since the classroom cluster level does not need to be specifically modeled to produce the unbiased random effects estimates of students and schools. However, if the classroom-level predictor W_{jk} is still of research interest and is modeled as a disaggregated at the student-level $W_{i(j)k}$, then the standard error estimates of its' coefficient still need to be adjusted by the square root of $1 + (n_L - 1)\rho$, as the clustering of higher school level still exists. Further, if $\rho = 0$, the cluster-level predictors $W_{i(jk)}$ and $Z_{i(jk)}$ (as shown in Eq. 2.7) do not need any clustering adjustments anymore. On this occasion, a single-level analysis using OLS estimation is sufficient as the data has a simple random sampling design.

When the estimated model is single-level OLS estimation, the variance inflation issue of the disaggregated classroom-level predictor $W_{i(jk)}$ is equivalent to the well-documented simple two-level clustering modeling situation where the teacher-level predictor W_j is modeled as $W_{i(j)}$. Consequently, the variance adjustment is constructed by the variance of the satisfactory two-level analysis which accounts for the clustering of students nested within classrooms, dividing

the variance of the single-level analysis. The variance estimates of the satisfactory two-level model is

$$\widetilde{Var}_{2L}(\gamma_{01}) = \left\{\sum_{j=1}^{J}(W_j{}'\,\boldsymbol{\psi_J} W_j)\right\}^{-1} = \sigma^2 \check{\tau}_{W_j}\left\{\sum_{j=1}^{J}(W_j{}'\,W_j)\right\}^{-1},$$

and $\boldsymbol{\psi_J}$ is the error variance-covariance structure

$$\boldsymbol{\psi_J} = \sigma^2\begin{bmatrix} 1 & \rho_0 & \cdots & \rho_0 \\ \rho_0 & 1 & \cdots & \rho_0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_0 & \rho_0 & \cdots & 1 \end{bmatrix}.$$

Similarly, the variance weighting index $\check{\tau}_{W_j} = 1 + (n_L - 1)\rho_0$. Finally, the variance inflation adjustment index $VOC_M^{(2-1,OLS)}$ for $W_{i(j)}$ in the single-level analysis using OLS estimation is the same as the two-stage *DEFF* (or *MF*). That is

$$VOC_M^{(2-1,OLS)} = \frac{\widetilde{Var}_{2L}(\gamma_{01})}{Var_{OLS}(\gamma_W)} = \check{\tau}_{W_j} = 1 + (n_L - 1)\rho_0. \qquad (2.9)$$

Obviously, fixing the teacher-level variance, $VOC_M^{(2-1,2L)}$ and $VOC_M^{(2-1,OLS)}$ increase as the average class size n_L increases. Therefore, the variance adjustment is more in need of models that are conducted for large class size contexts than the ones with small class size. Meanwhile, the number of classrooms in a sampled school (i.e., n_H) constraints in the practice setting of the potential value of n_L and ρ_0. This point is relevant in Chapter 3, in which an empirical example of omitting the middle cluster level is used to demonstrate the sensitivity analysis framework.

Furthermore, $VOC_M^{(2-1,OLS)}$ is smaller than $VOC_M^{(3-1,OLS)}$ by $(n_0 - 1)\rho_2$, which is intuitive as two sources of clustering dependency affect the estimation of the standard error estimate of the $Z_{i(jk)}$ coefficient estimate, while a single middle-level clustering affect the one of $W_{i(jk)}$. In other words, in the single-level analysis using OLS estimation, a Type I error issue could be more pronounced for the highest-level predictor than the middle level one.

Bias of the Standard Error Estimates of the Coefficients of $X_{i(j)k}$ and $X_{i(jk)}$

Finally, although the student-level predictor X_{ijk} is not the focus of the current study; its standard error estimate is upwardly biased when the clustering structure is omitted[8]. As evidenced in Moerbeek (2004)and Snijders (2005), the regression coefficient of an individual-level predictor X_{ij} in a two-level random intercept-only model tends to be upwardly biased when the adjacent upper cluster level is omitted in either the two- or single-level models. Type II error is also undesired since important individual-level predictor effects could be masked as insignificant. In a satisfactory random intercept two-level HLM model, the design effect formula of the standard error estimate of X_{ij}'s coefficient is $DEFF_2^{X_{ij}} = 1 - \rho_0$, which is less than 1 when $\rho_0 > 0$, indicating that the multi-stage sampling design is more efficient than the simple random sampling in this setting (Snijders, 2005). It is easy to extend to a three-level case for the variance estimate adjustment of the coefficient of $X_{i(jk)}$ from the OLS estimation case, which is:

$$VOC_M^{(1-1,OLS)} = 1 - \rho_1 = 1 - \rho_0 - \rho_2. \quad (2.10)$$

[8] When X_{ijk} is the predictor of interest while the cluster-level predictors and the random effects are not the foci, researchers could employ the fixed effect framework to account for the overall clustering dependency. However, when the cluster-level predictors are of the research interest, the fixed effect approach is less optimal. In the current setting of when the omitted cluster level data is not available, the shown design-based approach with the sensitivity analysis framework (in Chapter 3) could be preferred.

For the estimated two-level model case,

$$VOC_M^{(1-1,2L)} = \frac{1-\rho_0-\rho_2}{1-\rho} = 1 - \frac{\rho_0}{1-\rho}, \tag{2.11}$$

which is the ratio of the design effects of the satisfactory three-level model and the false two-level model. In Chapter 4, the main predictor of interest $W_{j(k)}$ encounters the same issue, in which a detailed derivation procedure is provided. Lastly, Table 2.3 below summarizes the VOCs of cluster-level predictors when omitting the teacher-level cluster only and omitting the clustering structure completely.

Table 2.2 A summary of VOCs when the middle cluster level is omitted in a three-level structured clustering data.

Three-level HLM		**Two-level HLM**			**Single-level OLS Estimation**		
Level	*Predictor*	*Level*	*Predictor*	*Variance adjustment*	*Level*	*Predictor*	*Variance adjustment*
Student	X_{ijk}	Student	$X_{i(j)k}$	$VOC_M^{(1-1,2L)} < 1$	Student	$X_{i(jk)}$	$VOC_M^{(1-1,OLS)} < 1$
Teacher	W_{jk}		$W_{i(j)k}$	$VOC_M^{(2-1,2L)} > 1$		$W_{i(jk)}$	$VOC_M^{(2-1,OLS)} > 1$
School	Z_k	School	$Z_{(j)k}$	$VOC_M^{(3-2,2L)} = 1$		$Z_{i(jk)}$	$VOC_M^{(3-1,OLS)} > 1$

Note. The letters in the parentheses of predictors' subscripts indicate the corresponding cluster levels that are omitted.

2.4.3 Simulation Results

A simulation study is designed to test the estimation bias when the middle cluster is omitted and the performance of the derived VOC formulas. In total, 12 conditions of random effect standardized variances and cluster sizes are set, and 500 replications are generated for each condition. The total sample size of students (M_I) and schools (M_K) are 5000 and 100, respectively, which fixes an average school size (i.e., the average number of sampled students

within a school) n_0 of 50. The setting of the average class size (i.e., the average number of sampled students within a class) n_L is set to be 5, 10, and 25. The corresponding average numbers of sampled classrooms or teachers in a school n_L 10, 5, and 2, and the ratio measure of the class and school sizes η are 0.08, 0.18, and 0.49.

Hedges and Hedberg (2007) provided a comprehensive list of ICCs for planning CRT based on the commonly used multi-stage sampled educational data sets, such as ECLSK and National Educational Longitudinal Study (NELS). They found that the ICCs are around 0.2 across all grades of all sample schools. Therefore, with setting $\sigma^2_{Total} = 1$, the values of random effects variance in the current study cover the conventional situations when the school-level random effects are relatively small (i.e., $\sigma^k = \rho_2 = 0.2$) and large (i.e., $\sigma^k = \rho_2 = 0.7$). Then, the teacher-level random effects of σ^j (= ρ_0) are 0.2, 0.5, and 0.7 to meet the conditions of equaling to, larger than, and smaller than the school-level random effects. Finally, the simulation study employed R package lme4 (Bates et al., 2015), where Restricted Maximum Likelihood (REML) is specified for estimating the variance component to accommodate the cases with small cluster samples.

The index of relative bias is computed to measure the magnitude of the estimation bias

$$\text{R. B.}_{est} = \frac{\tilde{\theta} - \theta}{\theta} = \frac{\tilde{\theta}}{\theta} - 1,$$

where θ represents the true parameters from the three-level model, including the random effects variances, standard errors of the teacher-level predictor W_{jk} , and the school-level predictor Z_k. Correspondingly, $\tilde{\theta}$ represents the estimates from the estimated two-level model or the disaggregated OLS estimation. Falsely estimated models lead R. B._{est} to deviate from zero.

Furthermore, a negative R. B.$_{est}$ represents underestimation and a positive R. B.$_{est}$ represents overestimation. Similarly, a relative bias index of R. B.$_{adj.est}$ is provided to show the need and performance of adjustments of estimates, in which $\tilde{\theta}$ becomes the ones that are adjusted by VOCs or the repartitioned variance-covariance formulas. The better performance of the adjustment of estimates, the closer to zero R. B.$_{adj.est}$ is. The simulation outputs are summarized in the following, and Appendix 2.C lists the parameter settings and provides detailed simulation results.

Bias of the Random Effects and the Adjustment Performance

The estimated two-level models overestimated the individual-level residual variance and school-level random effects variance, where the mean R. B.$_{est}$ are all positive and increasing with the increased σ^j or ρ_0. With increasing n_L and η, the magnitude of the overestimation of $\tilde{\sigma}^{(k)}$ increases while decreasing in $\tilde{\sigma}^{(i)}$. When the omitted between-classroom-within-school and the between-school variation only take 20% of the total variance respectively (i.e., $\sigma^j = \sigma^k = 0.2$) and the individual residual takes the most of the total variance, the overestimation of $\tilde{\sigma}^{(k)}$ is small, particularly when the average classroom size is relatively small (i.e., $n_L = 5\ or\ 10$) and the mean R. B.$_{est}$ of $\tilde{\sigma}^{(k)}$ is less than 0.01. Under the same conditions, however, the overestimation of the residual variance $\tilde{\sigma}^{(i)}$ is large, with $\tilde{\sigma}^{(i)}$ capables of being around three times as large as the true parameter. In an extreme converse case where the omitted between-classroom-within-school variation is considerably large (i.e., $\sigma^j = \rho_0 = 0.7$) and the individual variance and the between-school variation are small (i.e., $\sigma^i = 0.1$ and $\sigma^k = 0.2$), $\tilde{\sigma}^{(k)}$ can be as twice as large as the true parameter σ^k. The overestimation of $\tilde{\sigma}^{(i)}$ is extreme that $\tilde{\sigma}^{(i)}$ can be over seven times larger than the true parameter σ^i.

These patterns are consistent with the Eqs. 2.3–2.5 that $\tilde{\sigma}^{(i)}$ has a higher degree of overestimation compared with $\tilde{\sigma}^{(k)}$ under the same conditions of σ^{j} and n_L. Moreover, the adjusted variances performed considerably well, as the mean R. B.$_{adj.est}$ are close to 0 across all conditions.

Bias of the Standard Error Estimates of the Coefficients of $Z_{(j)k}$ and $Z_{i(jk)}$ and the Adjustment Performance

The absolute mean R. B.$_{est}$ of the standard error estimates of $Z_{(j)k}$ in the two-level models are all highly close to 0 (less than 0.01), which supports the previous derivation of $VIF_M^{(3-2)}$=1.In the single-level model using OLS estimation, the standard error estimates of $Z_{i(jk)}$ are consistently underestimated since the mean of R. B.$_{est}$ are all negative, and the standard deviations of R. B.$_{est}$ are nearly zero. The standard error estimates are only around 20 to 30 percent of the true parameter, which is relatively stable across all the conditions. This is because OLS estimation ignored the overall error clustering dependency so that distinguishing the sources of clustering matters less.

Bias of the Standard Error Estimates of the Coefficients of $W_{i(j)k}$ and $W_{i(jk)}$ and the Adjustment Performance

As shown by the negative value of the mean R. B.$_{est}$ and the nearly zero standard deviation of R. B.$_{est}$, the standard error estimates of the $W_{i(j)k}$ coefficient in the estimated two-level models and the $W_{i(jk)}$ coefficient in the OLS estimated single-level models is downwardly biased in all conditions.

In the two-level models, the standard error estimates are mostly underestimated when the omitted σ^j and n_L are large, and when σ^i is small. When $\sigma^j = 0.7$ and $\sigma^k = 0.2$, the standard error estimates can be a half and even only 20 percent of the parameter. When $\sigma^j = 0.2$ and $\sigma^k = 0.2$, the standard error estimates can still be only 40 to 70 percent of the parameter, which is non-trivial. Further, in the extreme case of when the individual residual variance is considerably small (e.g., $\sigma^i = 0.1$), the underestimation of standard error estimates is comparable in cases of either the majority clustering dependency coming from the school-level (i.e., when $\sigma^j = 0.2$ and $\sigma^k = 0.7$) or from the classroom-level (i.e., when $\sigma^j = 0.7$and $\sigma^k = 0.2$). This is intuitive from $VIF_M^{(2-1,2L)} = 1 + (n_L - 1)\rho_1$. These patterns are also found in the single-level models where the underestimation is positively related to the size of σ^j and n_L.

The performance of $VIF_M^{(2-1)}$ is generally good in almost all cases since $\text{R. B.}_{adj.est}$ has absolute mean and standard deviation values less than or around 0.1. However, one exception in the two-level models is when $n_L = 25$, $\sigma^j = 0.7$ and $\sigma^k = 0.2$ and the underestimation adjustment is not enough. The adjusted standard error estimate is around 75 percent of the true parameter, though having improved largely as compared with the unadjusted one of being 20 percent of the true parameter. In single-level models when $n_L = 5$, $\sigma^j = 0.2$ and $\sigma^k = 0.7$, the standard error estimates are over-corrected that the adjusted estimates are, on average, 20% larger than the parameter. In this case, the underestimation bias from the single-level model is close to zero (i.e., -0.05) that no adjustment is required in the first place.

Bias of the Standard Error Estimates of the Coefficients of $X_{i(j)k}$ and $X_{i(jk)}$ and the Adjustment Performance

Finally, the simulation found evidence of the overestimation bias of the standard error estimates of the coefficients of $X_{i(j)k}$ and $X_{i(jk)}$. This finding is consistent with Moerbeek (2004). Particularly in cases where σ^j and σ^k are large, the bias is substantial. When σ^j and σ^k are small as in $\sigma^j = 0.2$ and $\sigma^k = 0.2$, the R. B.$_{est}$ of the two-level HLM are less than 0.1. This resonates in Wang et al. (2019)'s simulation setting with small σ^j, σ^k, and corresponding evidence that shows that the standard error estimates *of* $X_{i(j)k}$ is unbiased.

2.5 Discussion and Conclusion

Extending an emerging body of research debating whether a middle cluster level matters in making the decision of using a two- or three-level model, this chapter summarizes and clarifies when a two-level model omitting the middle cluster level would impact the standard error estimate of a certain level predictors' regression coefficient in the settings of multi-stage sampling and CRT design. In previous studies, the relevant evidence is often shown through simulation and empirical analyses as examples. The current study complements those evidence by producing critical formulations of quantifying the standard error estimation bias (i.e., the correction index of VOCs), which are functions of the clustering parameters of the omitted middle cluster level. Simulation evidence is provided with settings of the practical K-12 education to aid for empirical implications.

Also, the findings shown by the VOCs formulas provide a general conclusion of the statistical mechanisms causing the bias and to what degree. The VOCs are specifically listed in

the above Table 2.2. For recommendations of modeling three- and two-level models, if the middle cluster level is a deliberate stage in sampling, even if this level is not directly related to the research questions, this cluster level should be explicitly modeled to correctly reflect the complete picture of the study designs of sampling stages and the levels of experimental mechanisms. An estimated two-level model omitting the middle cluster level should be corrected with the variance estimates of the random effects, whereas it would not produce biased standard error estimates of the coefficients of the third level predictors.

If the middle cluster level is incidental instead of being a deliberate sampling stage or receive treatment assignment, whether to model this level as random effects largely depend on whether the research interests relate to the predictors at this middle level. Many times, the middle cluster level conveys important mechanisms that researchers would prefer to include this middle level and corresponding predictors in the three-level models. Particularly, a two-level model in this situation would easily falsely disaggregate the middle-level predictors at the lowest level. In this case, the standard error estimates of the disaggregated middle-level predictors' coefficients need to be corrected to avoid Type I error.

This study also extends omitting the single middle cluster level to completely omitting the clustering of both the middle and highest levels. This extension contributes to the conventional design-based robust standard error studies, which do not distinguish the sources of dependencies in multilevel data structures while capturing the overall dependency. This point is best supported by the VOC derivation of the highest cluster level predictor. Additional to the omitted one cluster level scenario, this chapter also extends to the omitting the overall clustering dependency case

as the estimated model is a single-level model. Then, the cluster-level predictors estimates would have Type I error issues and the individual level predictor would have a Type II error. Moreover, the Type I error issue is more pronounced in the highest-level predictor than the middle-level one.

The above finding is empirical guidelines for researchers to decide whether the middle cluster level should be modeled. Further, combining with the sensitivity analysis framework and empirical examples presented in the following Chapter 3, researchers would further benefit from testing the magnitude of the robustness inference if a potential middle cluster level is not modeled. The current model-based design sets the basic random intercept model as the satisfactory model. If the random-slope model is the satisfactory model, the error variance-covariance matrix of Y_k and the standard error estimate expressions should be accommodated (see Snijders and Bosker, 1993). However, the random intercept model is a widely used model in education empirical research and an ideal starting point for more complex models in future research. Another limitation of this study is that the modeling setting assumes balanced designs, which is not always plausible in practice. Future work needs to develop VOCs, particularly in the CRT designs (Konstantopoulos, 2010), to accommodate unbalanced situations, such as when including the ratio of cluster sizes.

CHAPTER 3

SENSITIVITY ANALYSIS FRAMEWORK OF OMITTED CLUSTERING

3.1 Introduction

Good scientific research is expected to present the best design and models that can answer the research questions and satisfy the model assumptions. However, as argued earlier, the issue of omitting a cluster level in two-level HLM cannot be solved by a model-based approach (i.e., three-level HLM) in many practical situations, such as data restrictions and unidentifiable error variance-covariance structures. Given these concerns about omitted clustering, Chapter 2 (and later Chapters 4 and 5) provided formulas to quantify the standard error estimation bias of the coefficients, which are functions of the clustering parameters (i.e., ICCs and cluster sample size) of the omitted cluster levels. Further, the current chapter builds a sensitivity analysis using the VOCs to test the magnitude of the inference robustness when the model-based approach is not feasible. In practice, if empirical researchers aim to know how robust the inference they made from the estimated model with a potentially omitted cluster level, they may hypothesize the clustering parameters of the omitted cluster and utilize the this sensitivity analysis framework. .

In essence, the proposed sensitivity analysis evaluates the deviations of the estimated models from the ideal case of when all crucial random effects are correctly modeled and specified. Simply stated, the larger deviations from the assumption there are, the higher bias of the standard error estimates due to the omitted clustering, and the less robustness of the statistical inference. Panel (a) of Figure 3.1 demonstrates this idea. As defined earlier in Chapter 2, the satisfactory model is the ideal model that meets all the clustering assumptions, which is unknown

in practice and thus outlined with dashed lines in the figure on the right end. The estimated model is the actual conducted model and hypothesized with an omitted necessary cluster level, which could produce biased estimates. The more the estimated model deviates from the satisfactory model due to the omitted clustering, the less robust it is. The size of the deviation is then quantified by the hypothesized clustering effect via setting parameters of ICCs and cluster sizes of the omitted cluster level. Consequently, even if the satisfactory model is unknown, it can be hypothesized to test how far the estimated model deviates. In Figure 3.1, Model A and Model B are such two hypothesized satisfactory models. Specifically, the estimated model deviates from Model B farther than Model A, since Model B sets with larger clustering parameters.

The size of the deviation from the estimated model to a hypothesized satisfactory model can be represented in terms of the size of the bias of standard error estimates. Thus, if the deviation is considerable, the bias of the standard error estimates can be large enough to generate a false inference with either a Type I or Type II error. Therefore, a threshold satisfactory model defining the minimal deviation size to invalidate an inference is added in panel (b) of Figure 3.1. This idea is built on the "switch point" framework of Frank, Maroulis, et al. (2013), which defines a lower threshold of a non-zero effect study switches to a no effect one.

The clustering setting of the threshold satisfactory model is then the threshold clustering of the omitted cluster, which can help researchers quantify the robustness of their inferences to omitted clusters. For example, if researchers think Model A fundamentally represents the omitted clustering, then the estimated model does not produce a false null hypothesis decision since the threshold model is on the right of Model A. In this case, the estimated model would be acceptable, although its interpretations and implications should not be overstated. On the contrary, if Model B's clustering setting is also reasonable, the magnitude of the standard error

estimate bias of the estimated model is large enough that the estimated model generates a false decision of null hypotheses.

In Frank, Maroulis, et al. (2013), the *robustness* of inference is defined as "the evaluation of the estimate against the threshold (p. 439)". This definition constructs the amount of bias in an estimate to invalidate the inference. The threshold estimate is often specified as the one for statistical significance associated with an exact p-value of 0.05. Switch to the current study, the magnitude of the robustness of inference is evaluated by the deviation of the estimated model from the hypothesized satisfactory models. The larger the deviation, the less robust the inference regarding modeling specification on clustering dependency. Unlike a fixed threshold estimate in Frank, Maroulis, et al. (2013), the position of the hypothesized satisfactory model defined in the current study is flexible as shown in Figure 3.1, which changes along with the sizes of the omitted clustering degree (i.e., VOCs). The current study constructs a sensitivity measure accordingly: the percentage of reduced robustness of inference. This measure quantifies the magnitude of threats to the robustness of inference due to an omitted cluster level. The initial robustness of the estimated model should be considered 100% when the estimated model and the hypothesized satisfactory model has no distance (i.e., no omitted clustering issue). If the estimated model has an omitted clustering issue that a deviation between the estimated and the hypothesized satisfactory model exists, its initial robustness magnitude should be smaller than 100%. Thus, as the deviation increases, the robustness decreases.

Extend the sensitivity analysis application to treatment evaluation studies, the percentage of reduced effect size as a second sensitivity measure is developed. Further, when the hypothesized satisfactory model is on the right of the threshold model (such as the standard error associated with an exact p-value 0.05), a measure evaluates the risk of making a false null

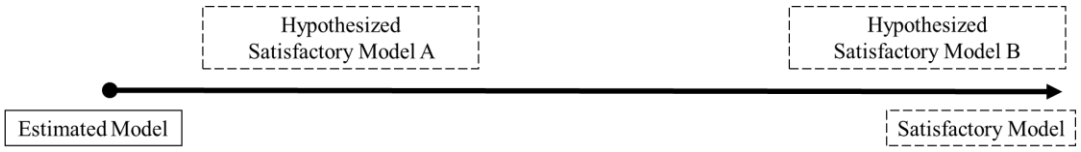
hypothesis decision is provided. For example, in the panel (b) of Figure 3.1, a red line presents the distance between the threshold satisfactory model and Model B. As the distance increases, the risk of making a Type I error (or a Type II error) increases. Further, the risk of having an invalid inference can be compared across hypothesized satisfactory models. The following discussion focuses on the scenarios of making Type I error, while Appendix 3.1 further provides the Type II error discussions.

Section 3.2 starts with the simple scenario of conducting a false single-level analysis which omits a higher cluster level and leads to underestimated standard error estimates. The developed measures and formulas are easily applied to the false two-level HLM with omitting a cluster level cases, and can also accommodate to the Type II error cases when the standard error estimates are upwardly biased (such as in the omitting highest cluster level case in Chapter 4). Section 3.3 provides an empirical example of employing the developed sensitivity analysis. The empirical example serves the discussion in Chapter 2, where a two-level HLM model is estimated while an incidental middle cluster level is potentially omitted.

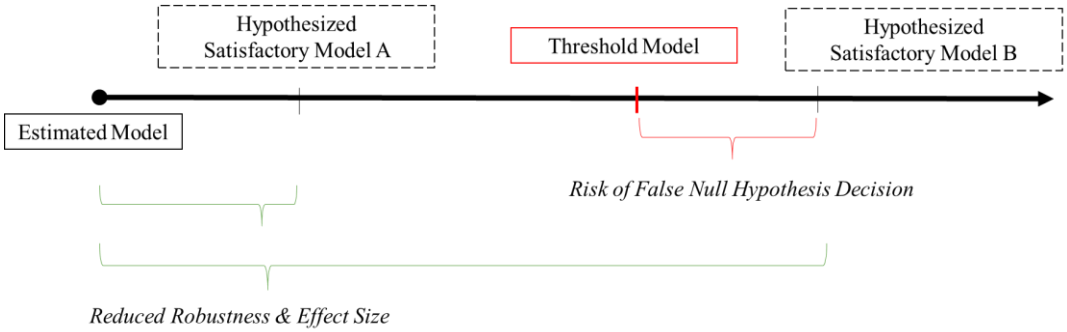
3.2 Constructing the Sensitivity Measures for Inference Robustness of Clustering

In Frank, Maroulis, et al. (2013), the magnitude of the inference robustness was quantified by constructing a ratio of a coefficient estimate with a threshold coefficient. Since the standard error estimate is of the focus of the current study in evaluating the impacts of the omitted clustering dependency, the current study construct the ratio of the t statistics from the estimated and hypothesized satisfactory models and the t critical value with an alpha level of 0.05, fixing the coefficient estimates if a cluster level is omitted (McNeish,2014). For example, consider an estimated single-level analysis with a continuous dependent variable $Y_{i(k)}$, which

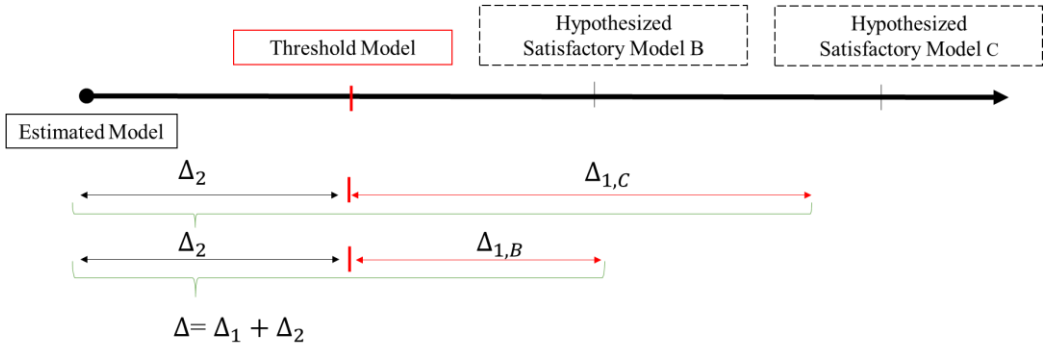
indicates the outcome of a student i in a school k though the school-level is omitted as shown in parenthesis: $Y_{i(k)} = \beta_{0(k)} + \beta_{1(k)}X_{i(k)} + \varepsilon_{i(k)}$. The coefficient estimate $\hat{\beta}_{1(k)}$of the predictor of interest $X_{i(k)}$ has a corresponding standard error estimate StE_{ols}.



(a) Deviation of the estimated model from the unknown satisfactory model



(b) Deviations of the estimated model from the hypothesized satisfactory models of A and B



(c) Deviations of the estimated model from the hypothesized satisfactory models of B and C

Figure 3. 1 Graphic demonstrations of the conceptualizing the sensitivity analysis framework

With the omission of the higher cluster level of schools, the standard error estimate is downwardly biased and needs adjustment, which turns to be $StE_{voc} = StE_{ols} * \sqrt{VOC}$, while the

point estimate $\hat{\beta}_{1(k)}$ remains the same. The VOC here is the design effect $1 + \rho_{icc} * (\overline{N}_k - 1)$, where the expected intraclass correlation ρ_{icc} and the average cluster size $\overline{N}_k$ are the clustering parameters. Further, setting the common 0.05 alpha level, the threshold model has the *t* critical value of $t^{\#} = 1.96$ and the standard error of $StE^{\#} = \hat{\beta}_{1(k)}/1.96 \cong \hat{\beta}_{1(k)}/2$. The following uses a general coefficient estimates notation $\hat{\beta}$ replacing $\hat{\beta}_{1(k)}$.

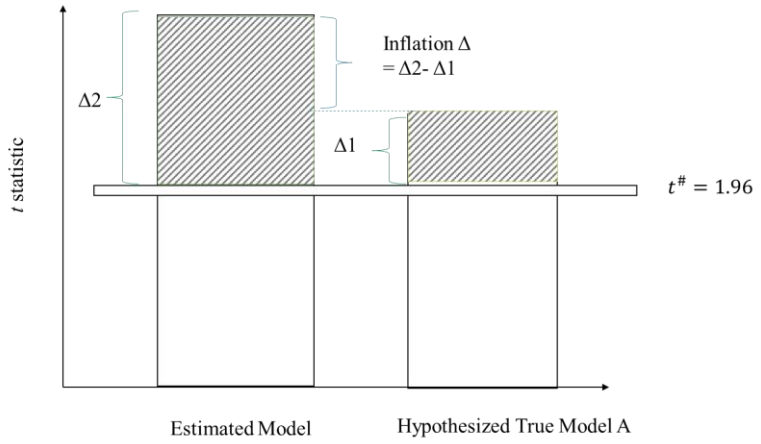
3.2.1 Scenario of No Type I error

This scenario presents the case of the estimated model (i.e., the single-level model using OLS estimation) deviating from the hypothesized satisfactory Model A with reduced inference robustness and effect size. However, the deviation is not large enough to result in a Type I error, as Model A is on the left of the threshold model. After transforming into the *t* statistic robustness framework as shown in Panel (a) Figure 3.2, this scenario yields $t_{ols} > t_{voc} > t^{\#}$, in which the *t* statistic from the estimated model is larger than the threshold $t^{\#}$ by Δ_2 (i.e., $\Delta_2 = t_{ols} - t^{\#}$), and the *t* statistic from Model A is larger than the threshold $t^{\#}$ by Δ_1 (i.e., $\Delta_1 = t_{voc} - t^{\#}$). The deviation of the estimated model and Model A is thus equivalent to the distance Δ between those two differences of *t* statistics against $t^{\#}$ (i.e., $\Delta = \Delta_2 - \Delta_1 \geq 0$).

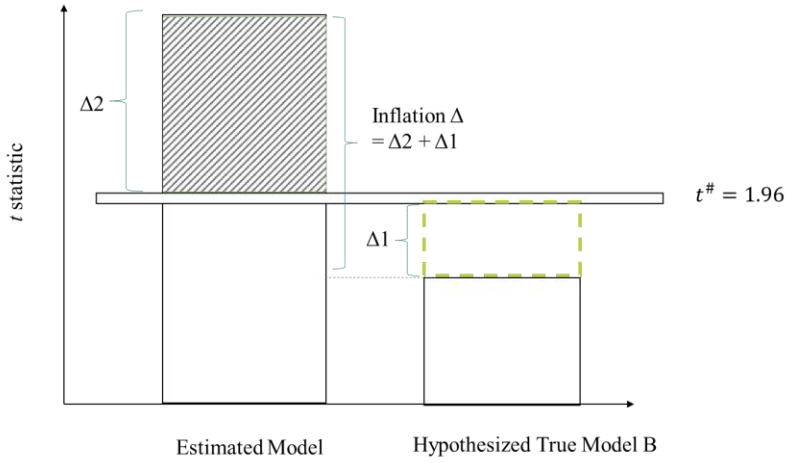
The larger the distance Δ, the larger inflation the *t* statistic of the estimated model is, and the stronger evidence of the reduced magnitude of robustness. Scaling Δ by t_{ols} as quantifying the size of inflation relatively to the t-statistic, the percentage of the reduced robustness is formulated as

$$W_{OC} = \frac{\Delta}{t_{ols}} = \frac{\Delta_2 - \Delta_1}{t_{ols}} = \frac{t_{ols} - t_{voc}}{t_{ols}} = 1 - \frac{StE_{ols}}{StE_{voc}} = 1 - \frac{1}{\sqrt{VOC}}. \qquad (3.1)$$

Consider Figure 3.2 Panel (a) below, when $\Delta = 0$ that $W_C = 0$, there is no bias in the standard error estimates due to potential omitted clustering. This is the case of the estimated model is the best practice model which initial robustness regarding with modeling clustering can be considered as 100%. With the increase of Δ, the initial robustness decreases by W_C.



(a) Scenario of No Type I error



(b) Scenario of having Type I error

Figure 3.2 Two Scenarios of Comparing *t* Statistics of the Estimated Model and the Hypothesized Models ($t_{ols} > t^{\#}$)

Further, I propose a measure of the changes in effect size. In educational research, particular in the experimental design research, the generic idea of effect size is the standardized

mean differences, which is the ratio of the treatment effect to a standard deviation (Hedges, 2007b). Then, the effect size of the predictor of interest $X_{i(k)}$ from the estimated single-level model is $ES_{OLS} = \frac{\widehat{\beta}_{1(k)}}{\sigma_{ols}}$, where the numerator is the fixed coefficient $\hat{\beta}_{1(k)}$ and the denominator is the standard deviation $\sigma_{ols} = StE_{ols} * \sqrt{N}$. And N is the total sample size. This definition of effect size is the adapted from Cohen's *d* (Cohen, 1962, 2009). Correspondingly, the effect size from the hypothesized satisfactory model[9] is $ES_{VIF} = \frac{\widehat{\beta}_{1(k)}}{\sigma_{voc}} = \frac{\widehat{\beta}_{1(k)}}{StE_{voc}*\sqrt{N}}$. Then, the percentage of the reduced effect size due to an omitted cluster level can be calculated using

$$ES_{OC} = \frac{ES_{OLS} - ES_{VOC}}{ES_{OLS}} = 1 - \frac{StE_{ols}}{StE_{voc}} = 1 - \frac{1}{\sqrt{VOC}}, \tag{3.2}$$

which is identical to W_C. As specified by the scenario setting, $\sqrt{VOC}$ here is smaller than the threshold $\sqrt{VOC_0}$ that the estimated model is acceptable as it does not lead to a false decision on a non-effective intervention or mechanism. However, the decisions made on the estimated effect size need to be cautious as the satisfactory effect size can be smaller.

In the context of education interventions and policy evaluations, there are several commonly used measures of interpreting effect size, such as the magnitude, cost of a program, and scalability of programs (Kraft, 2020). As a complement, ES_C can be considered as a sensitivity measure serving to quantify the uncertainty of effect size due to the omitted clustering effect. Noticeable, ES_C is different from the conventional sampling uncertainty measures of effect size, such as the standard error and confidence interval (see Cooper et al., 2019).

[9] The current effect size formula is constructed based on Cohen's *d*, while other definitions of effect size that satisfy specific research interest exist. A summary and comparison of commonly used effect size measures can be found in Fritz, Morris, & Richler (2012), and the ones developed for multilevel analysis can be seen in Hedges (2007).

Obviously, the size of ES_C depends on the values of the hypothesized clustering degree VOC and the original effect size estimate of the tested study. By hypothesizing meaningful settings of clustering degree (i.e., VOC and its parameters of ICC and cluster size) within the context of a certain study[10], ES_C constructs an interval as well. Then, multiplying the original effect size estimate with the range of ES_C, researchers gain an interval of effect size due to plausible omitted clustering settings. The larger the VOC, the larger reduction of effect size when fixing the original effect size. The wider range of the VOC, the more uncertainty of a study due to the omitted clustering.

When fixing the VOC, the same value ES_C could lead to different meanings with respect to different original effect size. For example, when $ES_C = 0.3$, a large effect size estimate of 0.3 only reduces to 0.2, which is still considerably large to indicate an effective and significant program. However, a medium effect size estimate of 0.1 reduces to 0.07, which would lead to a consideration of less strength of the detected effect. As shown, though a 3% reduction of a small effect size (i.e., 0.03 in the example) is much smaller than a 3% reduction of a larger effect (i.e., 0.1 in the example), the judgments on the reduced effect size realize on the magnitude of the original effect size. It is an advantage of ES_C measuring the percentage of reduction against the original effect size instead of being an arbitrary value of reduction. The interpretation of effect size depends largely on the research context (Hedges, 2008; C. J. Hill et al., 2008; Kraft, 2020). Though it is beyond the scope of this study to discuss the benchmarks of interpreting the magnitude of effect size, the current study suggests employing a summarized schema for

[10] See Korendijk, Moerbeek , et al. (2010)'s suggestions in assessing the ICC setting in educational research designs.

interpreting effect size along with the cost and scalability of programs from Kraft (2020, p. 20) when interpreting the magnitude of the reduced effect size of ES_C.

Researchers need to make decisions on setting plausible values of the clustering parameters of the omitted cluster when applying the above sensitivity analysis. In the setting of no Type I error, $\sqrt{VOC}$ is always smaller than the threshold $\sqrt{VOC_0}$. While, if $\sqrt{VOC}$ is possible to be larger than the threshold $\sqrt{VOC_0}$, the estimated model needs to further consider a Type I error issue discussed as following.

3.2.2 Scenario of Having a Type I error

A Type I error issue occurs when the estimated model is on the left of the threshold model while the hypothesized satisfactory model (i.e., Model B) is on the right, as shown in panel (b) of Figure 3.2, $t_{ols} > t^{\#} > t_{voc}$ when t_{VOC} is smaller than the threshold 1.96 by Δ_1 (i.e., $\Delta_1 = t^{\#} - t_{voc}$)while t_{ols} is larger than the threshold by Δ_2 (i.e., $\Delta_2 = t_{ols} - t^{\#}$). The estimated model deviates from Model B by $\Delta = \Delta_1 + \Delta_2$. The quantification process of the reduced robustness of inference and effect size is identical to the above scenario of no Type I error

$$W_{OC} = \frac{\Delta}{t_{ols}} = \frac{\Delta_2 + \Delta_1}{t_{ols}} = \frac{t_{ols} - t_{vif}}{t_{ols}} = 1 - \frac{StE_{ols}}{StE_{voc}} = 1 - \frac{1}{\sqrt{VOC}}, \tag{3.3}$$

$$ES_{OC} = \frac{ES_{OLS} - ES_{VOC}}{ES_{OLS}} = 1 - \frac{StE_{ols}}{StE_{voc}} = 1 - \frac{1}{\sqrt{VOC}}. \tag{3.4}$$

Further, as introduced earlier, a large distance between the threshold model and Model B suggests that the estimated model has a high possibility of making a Type I error. This Type I error risk can thus be quantified by the relative size of Δ_1 in Δ while fixing Δ_2

$$R_{OC} = \frac{\Delta_1}{\Delta} = \frac{\Delta_1}{\Delta_1 + \Delta_2} = \frac{1}{1 + \Delta_2/\Delta_1} = \frac{1}{1 + r}, \tag{3.5}$$

and

$$r = \Delta_2/\Delta_1 = \frac{t_{ols} - t^{\#}}{t^{\#} - t_{voc}} = \frac{\frac{\hat{\beta}}{StE_{ols}} - \frac{\hat{\beta}}{StE^{\#}}}{\frac{\hat{\beta}}{StE^{\#}} - \frac{\hat{\beta}}{StE_{voc}}}$$

$$= \frac{(StE^{\#} - StE_{ols})StE_{voc}}{StE_{ols}(StE_{voc} - StE^{\#})}$$

$$= \frac{(StE^{\#} - StE_{ols})\sqrt{VOC}}{StE_{ols}\sqrt{VIF} - StE^{\#}}, \tag{3.6}$$

where r is positive and $0 < R_{OC} < 1$ since Type I error only happens when $StE_{vif} > StE^{\#} > StE_{ols}$. In Panel (b) of Figure 3.2, fixing Δ_1 of the satisfactory model, a larger Δ_2 leads to larger risk of making the Type I error. This relationship is quantified through R_{OC} that the higher the omitted clustering effect or correspondingly the $\sqrt{VOC}$ is, the higher the risk it is of the estimated model for making a Type I error. Further, the value of R_{OC} makes comparisons with the threshold case of when $StE^{\#} = StE_{ols}$. This is because it is intuitive that when the satisfactory model has a *t* statistic that equals to the *t* threshold (that is $\sqrt{VOC} = \sqrt{VOC_0}$), the Type I error issue arises.

Back to Panel (c) of Figure 3.1, it further demonstrates how the risk index can be utilized for comparing hypothesized satisfactory models. A hypothesized satisfactory Model C has a higher clustering setting than Model B, and thus being located on the farther right of the threshold model than Model B, thus $\Delta_{1,C} > \Delta_{1,B}$. Also, fixing Δ_2, $R_{OC}^{C} > R_{OC}^{B}$. That is, if Model C is the satisfactory model, the estimated model has a higher risk of having a Type I error issue

than if Model B is satisfactory. Noticeably, since the relative size of Δ_2 in Δ is considered in the formulation, the ratio of R_{OC}^{C} and R_{OC}^{B} is not as simple as $\Delta_{1,C} > \Delta_{1,B}$. Researchers who intend to know the relative risks of having Type I errors across different clustering settings of VOC can further utilize a relative risk index of

$$R_d = \left| \frac{R_{OC}^{C} - R_{OC}^{B}}{R_{OC}^{B}} \right| = \left| \frac{R_{OC}^{C}}{R_{OC}^{B}} - 1 \right|. \tag{3.7}$$

In this manner, the risk of making Type I error increases by a percentage of R_d, if the omitted clustering setting of Model C is preferred than the one of Model B based on the research context. Finally, the above discussions focused on the Type I error issue. In Appendix 3.A, measures of robustness inferences are extended to the Type II error issue.

3.2.3 Heuristics Diagram and Interpretations of the Sensitivity Analysis

The heuristics diagram in Figure 3.3 depicts a possible flow of conducting the sensitivity analysis. Starting from the top of the diagram, researchers may first find the threshold $\sqrt{VIF_t}$. Solving Δ_1=0 (i.e., $t_{vif} = t^{\#}$), $\sqrt{VOC_t}$ yields

$$\sqrt{VOC_t} = \frac{\hat{\beta}}{1.96StE_{ols}} \cong \frac{\hat{\beta}}{2StE_{ols}} = \frac{1}{2} t_{ols}. \tag{3.8}$$

The use of this threshold $\sqrt{VOC_t}$ is straightforward, and it is of great use when empirical researchers need to anchor the threshold clustering parameters of the omitted cluster level. Further, researchers may set an empirical $\sqrt{VOC_0}$ with meaningful clustering parameter values of what best satisfies their prior knowledge about the suspected omitted cluster level. If the scientific $\sqrt{VOC_0}$ is unlikely to be exceeded at the threshold $\sqrt{VOC_t}$, then researchers may worry

less about the Type I error but focus on the magnitude of reduced robustness of inferences and effect size. If $\sqrt{VOC_0}$ exceeds the switch point value, then researchers need to further take into account the risk of having a Type I error. Setting a reasonable $\sqrt{VOC_0}$ value, researchers can manipulate the implications of an omitted cluster by exploring many possible values of the clustering parameters.

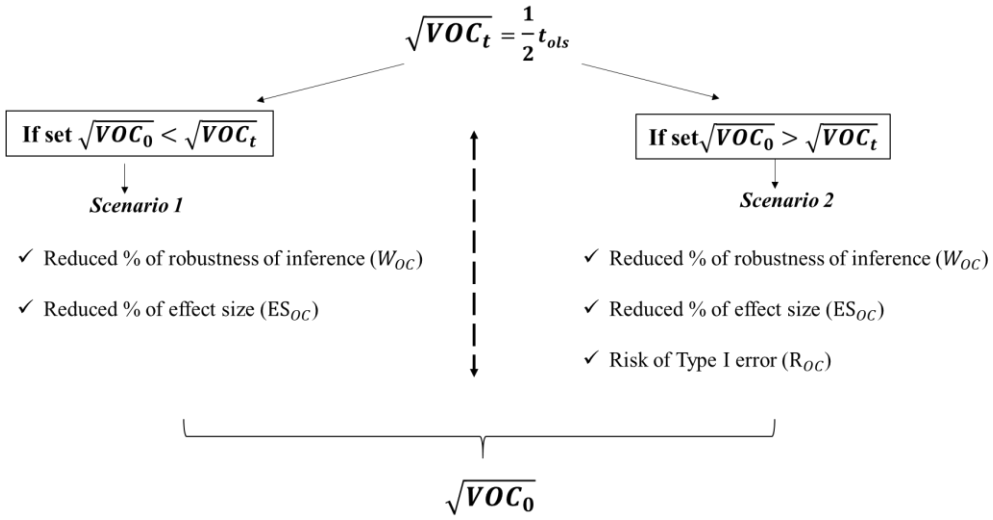


Figure 3. 3 Heuristics diagram of sensitivity analysis when the predictor of interest in the original single-level model is statistically significant.

Researchers can also conduct sensitivity analysis in the opposite direction. They may start with setting the clustering parameters to gain a $\sqrt{VOC_0}$, then judge with the $\sqrt{VOC_t}$. Enlightened by the work of Frank, Maroulis, et al. (2013), a sensitivity analysis can be of the most practical use by empirical research when it is equipped with a scientific language for interpretations. Here are the suggested interpretations of the above sensitivity analysis:

1) The robustness of inference (or effect size) reduces by *x* % (i.e., the values of W_{OC} or EF_{OC}) if the omitted cluster level has a clustering degree of *y* (i.e., the $\sqrt{VOC}$ value). The clustering degree is characterized by $\rho_{icc} = b$ and $\overline{N}_g = c$.

2) The risk of making Type I error increases by x % (i.e., the value of R_{OC}) if the omitted cluster level has a clustering degree of y.

3.3 Implication of the Sensitivity Analysis: Using an Empirical Example

This section provides an empirical study example to show how to use the sensitivity analysis in defining the robustness of inference when an incidental middle cluster level is omitted. The selected study is from Heafner et al. (2019), which examined demographic and course instruction related variables' impact on students' economics content knowledge. The employed data is the National Assessment of Educational Progress Economics Assessment (NAEP-E), which has a two-stage sampling design (with PSUs being schools and USUs being students). In that work, a two-level random intercept model is constructed, where the first and second levels are students and schools, because the authors mentioned that NAEP-E has data constraints to link students to teachers causing a three-level model to be prohibited (as seen in p. 331). In the final estimated model (see their Table 2 in p. 336), each level has corresponding demographic measures. Moreover, course type (such as AP course), curricular and instructional exposure (such as internet use in a class) measures are assigned at the student level.

It is reasonable to argue that some student-level predictors that are relevant to courses and instructions may be classroom-level predictors. For example, instructional exposure of reading in class and internet use for economic data may be uniform for students within the same classroom and teacher. Also, variations in the between-classrooms-within-schools cluster may be random. Therefore, the classroom level, as an incidental middle cluster level, is assumed to matter to be explicitly modeled.

The below sensitivity analysis, shown in Table 3.1, is performed to calculate the robustness of the inference of the student-level predictor of internet use for economic data. The statistics from the estimated models are presented in the section of estimated two-level HLM in the table, including the regression coefficient $\hat{\beta} = -1.44$ (I used the absolute value in the sensitivity analysis for simplicity reasons which does not affect the results), standard error estimates $StE_{2L} = 0.4$, the random effects variance of $\tilde{\sigma}^{(i)} = 22.62$ and $\tilde{\sigma}^{(k)} = 8.58$ conditioned on the predictors, the total number of the sample schools $M_K = 560$, and the average number of students within a school $n_0 = 20$. Meanwhile, the hypothesized average number of students within a classroom n_L and the between-classroom variance $\sigma^{(j)}$ need special attention since they together affect whether the VOCs and the corresponding calculated statistics of the three-level model (such as the random effects variance $\sigma^{(i)}$ and $\sigma^{(k)}$) are plausible. In Table 3.1, three values of n_L are hypothesized to provide cases of extreme small cluster size of classrooms and the regular ones.

Following the steps shown in the heuristics diagram of Figure 3.3, I first find the threshold $\sqrt{VOC_t} = 1.837$. This threshold is then used to calculate the corresponding ρ_0 and $\sigma^{(j)}$. In the cases of when n_L are 2 and 10, the threshold-based ρ_0 is not plausible since they exceed the boundary of (0,1). In these two cases, it is more meaningful to find the possible maximum and minimum ρ_0. For example, when $n_L = 2$, the maximum value of a ρ_0 is 0.665 to make the regression estimates in the hypothesized three-level HLM feasible. Further, even when ρ_0 is large, the corresponding $\sqrt{VOC_{max}}$ would not lead to a StE_{vif} that is larger than $StE^{\#}$. Thus, there is no need to concern about potential Type I error issue when the average classroom size is extremely small. However, the robustness of inference (or effect size) reduces by around

50 % (i.e., the values of W_{OC} or EF_{OC}), which is not trivial. These settings reflect the earlier discussion of no Type I error scenario in Section 3.1.1. The following shows the having Type I error scenario.

In the setting of $n_L = 10$, a minimum ρ_0 is needed to specify eligible regression estimates in the hypothesized three-level HLM. This $\rho_{0,min}$ is extremely small, being 0.01, which still can lead to a Type I error since the corresponding StE_{vif} is larger than $StE^{\#}$. The risk of making a Type I error (i.e., R_{OC}) increases by 0.02, compared with the threshold setting with the *t* statistic at the switch point of 1.96. Also, when $n_L = 10$, the feasible $\rho_{0,max}$ is 0.58 with a R_{OC} the maximum value of 0.24. Finally, when $n_L = 7$, the threshold-based ρ_0 is plausible for being 0.176, which means that any ρ_0 that is larger than 0.176 could result in Type I error or not if ρ_0 is smaller than 0.176. Two values of ρ_0 being 0.5 and 0.1 are used to demonstrate this point.

This section went through the implication of the sensitivity analysis framework. As shown by the above example that inferring the magnitude of the robustness inference largely depends on the selection of the clustering parameters of the omitted cluster level. In practice, researchers may require meaningful clustering parameters from the previous research evidence to make the best argument for the inference robustness. As shown in the above specific example, the calculated between-classroom variation as measured by $\sigma^{(j)}$ and ρ_0 are regulated by VOC formulas and empirical evidence. This evidence encourages researchers to be cautious about excluding the classroom-level in modeling and assign the classroom-level predictors to other levels.

Table 3.1 Sensitivity analysis of the student-level predictor: Internet-use for economic data

Estimated Two-level HLM						**Hypothesized Three-level HLM**							
				M_K	560	n_L	2		7			10	
				n_0	20	η	0.053		0.316			0.474	
$\hat{\beta}$	|-1.44|			$\tilde{\sigma}^{(i)}$	22.62	$\sigma^{(i)}$	20.748	19.812	5.499	15.600	19.500	0.312	4.524
$StE^{\#}$	0.735	$t^{\#}$	1.96	/	/	$\sigma^{(j)}$	2.964	2.964	17.121	11.946	3.120	22.308	18.096
StE_{2L}	0.400	t_{2L}	3.60	$\tilde{\sigma}^{(k)}$	8.58	$\sigma^{(k)}$	7.488	8.424	8.580	3.654	8.580	8.580	0.008
				ρ	0.275	ρ_2	0.240	0.270	0.219	0.117	0.275	0.270	<0.001
						ρ_0	0.665	0.095	0.176	0.500	0.1	0.010	0.580

$\boldsymbol{\eta}$	$\boldsymbol{n_L}$	$\boldsymbol{\rho_0}$		$\sqrt{\boldsymbol{VOC}}$		$\boldsymbol{StE_{voc}}$	$\boldsymbol{W_{OC}}$ & $\boldsymbol{ES_{OC}}$	$\boldsymbol{R_{OC}}$
				Threshold $\sqrt{VOC_t}$	1.837	0.735	0.456	*Switch Point*
0.053	2	$\sqrt{VOC_t}$ based ρ_0	2.215		*NA*	*NA*	*NA*	*NA*
		$\rho_{0,max}$	0.665	$\sqrt{VOC_{max}}$	1.380	0.552	0.275	*NA*
		$\rho_{0,min}$	0.095	$\sqrt{VOC_{min}}$	1.168	0.467	0.144	*NA*
0.316	7	$\sqrt{VOC_t}$ based ρ_0	0.176	$\sqrt{VOC_t}$	1.837	0.735	0.456	*Switch Point*
		ρ_0	0.5	$\sqrt{VOC}$	2.169	0.867	0.539	0.15
		ρ_0	0.1	$\sqrt{VOC}$	1.749	0.700	0.428	*NA*
0.474	10	$\sqrt{VOC_t}$ based ρ_0	-0.021		*NA*	*NA*	*NA*	*NA*
		$\rho_{0,min}$	0.01	$\sqrt{VOC_{min}}$	1.877	0.751	0.467	0.02
		$\rho_{0,max}$	0.58	$\sqrt{VOC_{max}}$	2.494	0.998	0.599	0.24

CHAPTER 4

OMITTED HIGHEST CLUSTER LEVEL

4.1 Introduction

The context of schools and districts play important roles in many aspects of education, which has been a major topic in educational effectiveness studies since the renowned "Coleman report" of the 1960s (Gamoran et al., 2000; Rumberger & Palardy, 2004). In many aspects, schools and districts provide particular social contexts, physical resources, and leadership distributions and provoke varying students learning outcomes (Akerlof & Kranton, 2002; Fahle & Reardon, 2018; Muijs, 2020; Muller, 2015; Xia et al., 2020). Current educational database, such as the NCES-initiated survey programs, provide many significant instruments measuring the contexts of schools and districts, as well as within-school and -district variations (Muller, 2015). Methodologically, if this rich contextual information is omitted in modeling, studies may give spurious conclusions since the satisfactory but omitted between-school (or district) variation would be trapped into the lower levels of classrooms and teachers, whose impacts would thus be falsely enlarged on students' learning (see Moerbeek, 2004, and other studies mentioned in Chapter 1).

This chapter intends to address the analytical issues of omitting a highest cluster level (such as schools and districts) in a two-level HLM model. Specifically, this chapter sets a conceptual two-level random intercept model examining students' learning outcome with school level predictors and assuming that an even higher cluster level of districts is omitted. Following Chapter 2's discussion on omitting the middle cluster level, this chapter also applies the mechanisms of sampling and experimental designs to the discussion of omitting a necessary

highest cluster level, which facilitate to answer *when* the highest clustering dependency matters in modeling. Popular educational survey data sets are used as examples for empirical concerns. Then, the question of how much the omitted cluster level matters in making a robust inference is answered by the derived VOC formulas and evidenced by a simulation study. Further, an empirical study example using two-level model is provided to implement the VOCs within the sensitivity analysis framework developed in Chapter 3.

4.2 Omitted Highest Cluster Level in Sampling and Experimental Design

4.2.1 Omitting PSUs in a Three-Stage Sampling Structure Data

PSUs could be omitted in empirical analysis with data that has a three-stage sampling design. For instance, the public available version date sets (e.g., ECLSK) are often do not provide linkable ID of SSUs of schools to PSUs of districts or counties[11]. In this case, two-level HLM models leave out the clustering of schools within districts or counties, although the clustering dependency due to students-nesting-within-schools is modeled explicitly. The design effect of the true three-stage sampling is

$$DEFF_{L3} = 1 + n_{(\mathrm{s2})}(n_{(\mathrm{s1})} - 1)\rho_{(\mathrm{s1})} + \left(n_{(\mathrm{s2})} - 1\right)\rho_{(\mathrm{s2})},$$

where $\rho_{(\mathrm{s1})}$ is the expected correlation among SSUs within a PSU, and $n_{(\mathrm{s1})}$ is the sample number of SSUs within a PSU. Also, $\rho_{(\mathrm{s2})}$ is the expected correlation among FSUs within an

[11] Sometimes, ignoring a sampling stage could happen to when the sampling scheme is not universal in a large survey study. For example, in some international survey programs, countries may vary in sampling scheme to accommodate local context. Researchers may easily use the general sampling scheme as the universal design while ignore certain exceptions. Chen and Rust (2017) introduced such a scenario in the Programme for International Student Assessment (PISA) 2015, which used a general two-stage sampling design where the two stages are schools and students (OECD, 2015). While PISA of Russia used a three-stage sampling design, where geographical areas are PSUs, schools are SSUs, and students are USUs (OECD, 2015). The PSUs of geographical areas maybe easily ignored if a two-stage sampling scheme is taken as universal when the research data employed is PISA of Russia.

SSU, and $n_{(s2)}$ is the sample number of FSUs within an SSU. With only one layer of clustering accounted from the second stage sampling, the corresponding design effect is measured by $\rho^*_{(s2)}$ and $n^*_{(s2)}$ as $DEFF^*_{2L} = 1 + \left(n^*_{(s2)} - 1\right)\rho^*_{(s2)}$.

Figure 4.1 below visualizes the structures of these two design effects. Obviously, when the first-stage sampling is omitted, $\rho_{(s1)}$ turns to be 0, since SSUs are now falsely assumed to be independent to each other even if they are in the same PSU. Therefore, $DEFF^*_{2L}$ is not sufficient in two ways. One is that the two distinct sources of clustering measured by $\rho_{(s1)}$ and $\rho_{(s2)}$ are now absorbed by a single clustering dependency (i.e., $\rho^*_{(s2)}$). The other one is that the sampling structure is reduced from $n_{(s1)} + n_{(s2)}$ to $n_{(s2)}$. Immediately, the $DEFF^*_{2L}$ overestimate the standard error of the estimate. This is because the effective sample size calculated based on $DEFF^*_{2L}$ is smaller than the true effective sample size given by $DEFF_{3L}$. Equivalent to the design-based approach, conducting a two-level HLM model with a three-stage sampling design, the omitted highest cluster level results in the repartitioned random effects and a shrinking error variance structure. The comparison of the design effects resonates with Moerbeek (2004) and Opdenakker and Van Damme (2000) which provided simulation evidence that omitting the highest cluster level results in inflated standard errors of the adjacent lower-level predictors' coefficient standard error estimates, and thus Type II errors. Later sections provide detailed mathematical procedures of formulating the biased standard error estimates.

PSU (District)			1	1	1	1	2	2	2	2	3	3	3	3
	SSU (School)		1	1	2	2	3	3	4	4	5	5	6	6
		USU (Student)	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	$\rho_{(s2)}$	$\rho_{(s1)}$	$\rho_{(s1)}$	0	0	0	0	0	0	0	0
1	1	2	$\rho_{(s2)}$	1	$\rho_{(s1)}$	$\rho_{(s1)}$	0	0	0	0	0	0	0	0
1	2	3	$\rho_{(s1)}$	$\rho_{(s1)}$	1	$\rho_{(s2)}$	0	0	0	0	0	0	0	0
1	2	4	$\rho_{(s1)}$	$\rho_{(s1)}$	$\rho_{(s2)}$	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	$\rho_{(s2)}$	$\rho_{(s1)}$	$\rho_{(s1)}$	0	0	0	0
2	3	6	0	0	0	0	$\rho_{(s2)}$	1	$\rho_{(s1)}$	$\rho_{(s1)}$	0	0	0	0
2	4	7	0	0	0	0	$\rho_{(s1)}$	$\rho_{(s1)}$	1	$\rho_{(s2)}$	0	0	0	0
2	4	8	0	0	0	0	$\rho_{(s1)}$	$\rho_{(s1)}$	$\rho_{(s2)}$	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	$\rho_{(s2)}$	$\rho_{(s1)}$	$\rho_{(s1)}$
3	5	10	0	0	0	0	0	0	0	0	$\rho_{(s2)}$	1	$\rho_{(s1)}$	$\rho_{(s1)}$
3	6	11	0	0	0	0	0	0	0	0	$\rho_{(s1)}$	$\rho_{(s1)}$	1	$\rho_{(s2)}$
3	6	12	0	0	0	0	0	0	0	0	$\rho_{(s1)}$	$\rho_{(s1)}$	$\rho_{(s2)}$	1

PSU (District)			1	1	1	1	2	2	2	2	3	3	3	3
	SSU (School)		1	1	2	2	3	3	4	4	5	5	6	6
		USU (Student)	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1		0	0	0	0	0	0	0	0	0	0
1	1	2	$\rho^*_{(s1)}$	1	0	0	0	0	0	0	0	0	0	0
1	2	3	0	0	1	$\rho^*_{(s1)}$	0	0	0	0	0	0	0	0
1	2	4	0	0	$\rho^*_{(s1)}$	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	$\rho^*_{(s1)}$	0	0	0	0	0	0
2	3	6	0	0	0	0	$\rho^*_{(s1)}$	1	0	0	0	0	0	0
2	4	7	0	0	0	0	0	0	1	$\rho^*_{(s1)}$	0	0	0	0
2	4	8	0	0	0	0	0	0	$\rho^*_{(s1)}$	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	$\rho^*_{(s1)}$	0	0
3	5	10	0	0	0	0	0	0	0	0	$\rho^*_{(s1)}$	1	0	0
3	6	11	0	0	0	0	0	0	0	0	0	0	1	$\rho^*_{(s1)}$
3	6	12	0	0	0	0	0	0	0	0	0	0	$\rho^*_{(s1)}$	1

Figure 4.1 Data correlation structures of three-stage sampling designs when the first sampling stage in is included and omitted.

4.2.2 Incidental Highest Level above PSUs

Many times, a higher cluster level emerges even if it is not designed in sampling but matters to answer the research questions. McNeish and Wentzel (2016) defined such highest cluster level as incidental level to distinguish from the deliberate levels of the sampling stages; they also provided several example scenarios of when such incidental highest cluster level would occur. One is that individual two-level data are integrated into a single data set to invest the studies' generalizability and power. This scenario applies to meta-analysis where individual studies' effect size estimates are combined to obtain a summary statistic in which effect sizes are nested within studies. Further, the studies are nested within investigators. Thus, the investigators

form an incidental highest cluster level, and the between-investigator variation could be relevant to the research question (Konstantopoulos, 2011).

Another scenario is that when certain large sample size of PSUs of schools is required, a relatively large sample size of districts will be incidentally presented as a higher cluster level, though it may not directly relate to the research questions. For example, the Education Longitudinal Study of 2002 (ELS:2002) has a two-stage sampling design where schools are PSUs and students are USUs (Stapleton & Kang, 2018). With 16,197 sampled schools nationwide in ELS, the districts level, with a considerably large sample size, is introduced naturally while the linked ID of schools and districts is not accessible in the public-use file. Hence, the district cluster level is omitted due to data restrictions.

The above examples require three-level models to account for the random variation at the incidental highest cluster level, particularly when the highest-level units are samples and the inferences are made to the population. Conversely, the incidental cluster level does not need to be included with random effect when they are population units. Take the study of Wong and Li (2008) as an example, which utilized a two-level model to examine school-level contextual factors' impacts on teachers' information and communication technology implementation effectiveness. As they stated that the sampled schools are from all 18 districts in the studied area, the districts are not required to be modeled as random. Similarly, the two-stage design approach with sampling design effect $DEFF_{2L} = 1 + \left(n_{(\mathrm{s1})} - 1\right)\rho_{(\mathrm{s1})}$ is adequate for the clustering dependency due to sampling. In this case, based on the estimated two-level model, a fixed effect framework can be further utilized for the higher-level districts (i.e., add dummy variables indicating memberships of districts) (McNeish & Kelley, 2019).

4.2.3 Omitted PSUs above the Level of Treatment Assignment

Consider a two-level model being conducted in a study where the outcome is at the individual student level and the assignment of treatment is at the higher school level. If the utilized data is a two-stage sampling design where PSUs and USUs are schools and students respectively, and the statistical inference aims to the population of schools, the estimated two-level model is appropriate to capture the clustering with the school-level random effect. This model is a typical CRT that has shown in Chapter 2.

Now consider a three-stage sampling structure data where PSUs are districts, SSUs are schools, and USUs are students. The above two-level model is no longer sufficient because the random effects of the highest cluster level is omitted. Furthermore, the CRT model turns to be a Block Randomized Trial (BRT) since the schools within districts are randomly assigned with treatments. The conceptual differences of these two designs are depicted in Figure 4.2. If the true PSUs of districts are omitted or hidden (as shown by the dashed ovals below the dashed line), the experimental design can be falsely interpretated as the treatments being assigned to the schools randomly and all students in each school received with the same treatment. With the presence of districts, schools within the blocks of districts are randomly assigned with treatments. Schools remain as clusters since students in each school received with the same treatment. See Hedges and Rhoads (2010) for a summary of the relationships between BRT and CRT.

Since the inference targets the population of districts and schools, the three-level BRT model explicitly models the between-district variation with the random effect of districts. Conceptually, the clustering dependency due to sampling is now sufficiently captured in addition to the clustering of assignment, whereas the (false) two-level CRT only models the latter source

of clustering dependency[12]. This argument is consistent with Abadie et al. (2017) that clustering is due to the distinct rationales of sampling and assignment.

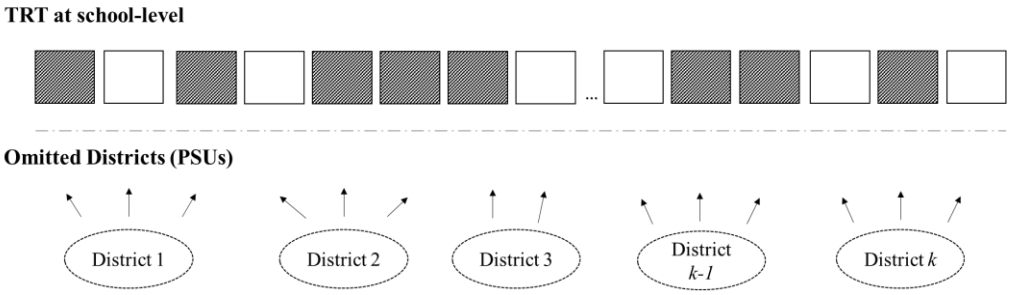


Figure 4.2 Omitted highest cluster level in a two-level CRT design

In the experimental design planning work of Hedges and Hedberg (2014), defining design parameters, such as ICCs, need to consider the omission of the districts as blocks while only keeping the schools as clusters. In such cases, the between-district variation is pooled into the between-school variation and the school-level ICCs are larger than they should be (Hedges & Hedberg, 2014, p. 455). Still, the effects on standard error estimates when omitting the highest cluster level in experimental design has not yet been extensively studied. Particularly, practical guidelines lack for empirical studies.

[12] Often, a three-level BRT model includes the random effect of the interaction term of treatment by district since the treatment effects' variation could depend on schools (see Konstantopoulos, 2008a, 2008b). The current paper does not include the random slope of the treatment and the corresponding interaction term in the later modeling settings in Section 4.2 to keep consistent with the setting of random intercept model of the whole study.

4.3 Quantification of Standard Error Bias

4.3.1 Model Setting

Follow the examples made above that the district cluster level is omitted, I first consider an estimated two-level random intercept model which only captures the clustering dependency of students (denoted as *i*) nested within a school *j*:

$$\text{Student-level: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \tilde{\varepsilon}_{ij},$$

$$\text{School-level: } \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + \gamma_{02}Z_j + \tilde{r}_{0j},$$

$$\beta_{1j} = \gamma_{10},$$

$$\text{Mixed model: } Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{02}Z_j + \tilde{r}_{0j} + \tilde{\varepsilon}_{ij},$$

where X_{ij} and W_j are student- and school-level predictors, and Z_j is modeled at the school level whereas it is truly a district-level measure. Also, predictors are group-mean centered so that the exogeneity assumption holds. In the setting of a two-level CRT design, W_j can be the binary treatment variable. The random effects of $\tilde{\varepsilon}_{ij}$ and $\tilde{r}_{0j}$ are assumed to be normally distributed with zero means, and variances of $\tilde{\sigma}^{(i)}$ and $\tilde{\sigma}^{(j)}$ respectively: $\tilde{\varepsilon}_{ij} \sim N\left(0, \tilde{\sigma}^{(i)}\right)$, $\tilde{r}_{0j} \sim N\left(0, \tilde{\sigma}^{(j)}\right)$, and $cov(\tilde{\varepsilon}_{ij}, \tilde{r}_{0j}) = 0$.

Identical to Chapter 2, for each school *j* from the total M_J sample schools, the error variance-covariance matrix of Y_j, denoted as $\tilde{\boldsymbol{\psi}}_{\boldsymbol{J}}$, is

$$\tilde{\boldsymbol{\psi}}_{\boldsymbol{J}} = var\left(Y_j\right) = \tilde{\boldsymbol{R}} + \boldsymbol{l}_{n_L}\tilde{\boldsymbol{G}}\boldsymbol{l}'_{n_L},$$

where n_L is the average school size (i.e., the average number of students within a school) and $\boldsymbol{l}_{n_L}$ is a column vector of n_L ones. There are M_J sample schools and the total sample size of students is thus $M_J * n_L$. The matrix $\widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{G}}$ reflect the composition of variance components at the student- and school level respectively:

$$\widetilde{\boldsymbol{R}} = \tilde{\sigma}^{(i)}\boldsymbol{I} = \begin{bmatrix} \tilde{\sigma}^{(i)} & 0 & \cdots & 0 \\ 0 & \tilde{\sigma}^{(i)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\sigma}^{(i)} \end{bmatrix}$$

and

$$\boldsymbol{l}_{n_L}\widetilde{\boldsymbol{G}}\boldsymbol{l}'_{n_L} = \begin{bmatrix} \tilde{\sigma}^{(j)} & \tilde{\sigma}^{(j)} & \cdots & \tilde{\sigma}^{(j)} \\ \tilde{\sigma}^{(j)} & \tilde{\sigma}^{(j)} & \cdots & \tilde{\sigma}^{(j)} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\sigma}^{(j)} & \tilde{\sigma}^{(j)} & \cdots & \tilde{\sigma}^{(j)} \end{bmatrix}.$$

Then,

$$\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}} = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} = \sigma^2\left[(1-\rho)\boldsymbol{I} + \rho\, \boldsymbol{l}_{n_L}\boldsymbol{l}'_{n_L}\right]. \tag{4.1}$$

The ICC $\rho = \frac{\tilde{\sigma}^{(j)}}{\sigma^2} = corr\left(y_{ij}, y_{i'j}\right)$ measures the expected correlations among any two randomly selected students from the same school. Now consider the satisfactory three-level random intercept model which includes the omitted highest level of districts (noted as *k*):

$$\text{Student-level: } Y_{ijk} = \pi_{0jk} + \pi_{1jk}X_{ijk} + \varepsilon_{ijk},$$

$$\text{School-level: } \pi_{0jk} = \beta_{00k} + \beta_{01k}W_{jk} + r_{0jk},$$

$$\pi_{1jk} = \beta_{10k},$$

District-level: $\beta_{00k} = \gamma_{000} + \gamma_{001} Z_k + u_{00k}$,

$$\beta_{01k} = \gamma_{010},$$

$$\beta_{10k} = \gamma_{100},$$

Mixed model: $Y_{ijk} = \gamma_{000} + \gamma_{100} X_{ijk} + \gamma_{010} W_{jk} + \gamma_{001} Z_k + u_{00k} + r_{0jk} + \varepsilon_{ijk}$.

The previously disaggregated predictor Z_j is now defined at the correct level of district as Z_k. Further, the random effect of the district-level is explicitly modeled and is assumed to be normally distributed with mean zero and variance of $\sigma^{(k)}$. Also, the random effects of the student- and school-level are assumed to have normal distributions, which have means of zero and variance of $\sigma^{(i)}$ and $\sigma^{(j)}$, respectively as $\varepsilon_{ijk} \sim N\left(0, \sigma^{(i)}\right)$, $r_{0jk} \sim N\left(0, \sigma^{(j)}\right)$, and $u_{00k} \sim N\left(0, \sigma^{(k)}\right)$. These random effects are independent to each other.

The three-level model has two ICCs, including the expected correlation among students within the same school and the same district $\rho_1 = \frac{\sigma^{(j)} + \sigma^{(k)}}{\sigma^2}$, and the expected correlation among students within the same district while from different schools $\rho_2 = \frac{\sigma^{(k)}}{\sigma^2}$. The average district sample size (i.e., average number of schools in a district) is n_H. Also, the total sample districts $M_K = \frac{M_J}{n_H}$, and the average number of students in a district is $n_H * n_L$.

The error variance covariance matrix $\boldsymbol{\psi}_K$ of a district *k* is

$$\boldsymbol{\psi}_K = \sigma^2 \{ \boldsymbol{I}_{n_H} \otimes \left[(1 - \rho_1) \boldsymbol{I}_{n_L} + (\rho_1 - \rho_2) \boldsymbol{l}_{n_L} \boldsymbol{l}_{n_L}{}' \right] + \rho_2 \, \boldsymbol{l}_{n_H * n_L} \boldsymbol{l}_{n_H * n_L}{}' \}, \qquad (4.2)$$

where $\boldsymbol{I}_{n_H}$ is a diagonal matrix with a dimension of the average cluster size of level-3 (*K*) $n_H * n_H$, $\boldsymbol{I}_{n_L}$ is a diagonal matrix with a dimension of the average cluster size of level- 2 (*J*) $n_L * n_L$,

$\boldsymbol{l}_{n_L}$ is a vector column of n_L ones, and $\boldsymbol{l}_{n_H * n_L}$ is a vector column of $n_H * n_L$ ones. Conceptually, $0 \leq \rho_2 \leq \rho$ and $1 \leq n_H \leq M_J$. I also define $\rho_0 = \frac{\sigma^{(j)}}{\sigma^2}$, which is the proportion of the true between-school variance in the total error variance, and ρ_0 is smaller than ρ by ρ_2. The detailed definition rationales of these ICCs have already been given in Chapter 2.

Figure 4.3 demonstrates the error variance-covariance structure of $\boldsymbol{\psi}_K$ from the three-level model and $\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}}$ from the two-level model omitting the highest cluster level of districts. Noticeably**,** compared with $\boldsymbol{\psi}_K$, the error correlation structure of $\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}}$ shrank from the $(n_H * n_L) * (n_H * n_L)$ block diagonal matrices (i.e., the purple dashed boxes) to the $n_L * n_L$ diagonal matrices (i.e., the orange highlighted squares). The shadowed areas represent the shrank segments due to falsely assumed independence among schools within districts.

When the highest cluster level is omitted, the between-district variation is fully redistributed to the between-school variation while the between-student variation remains the same, which are

$$\tilde{\sigma}^{(i)} \cong \sigma^{(i)}, \tag{4.3}$$

and

$$\tilde{\sigma}^{(j)} \cong \sigma^{(j)} + \sigma^{(k)}. \tag{4.4}$$

Then, $\rho = \frac{\tilde{\sigma}^{(j)}}{\sigma^2} = \frac{\sigma^{(j)} + \sigma^{(k)}}{\sigma^2} = \rho_0 + \rho_2 = \rho_1$. The shrank parts in $\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}}$ are ρ_0, which are falsely captured by ρ in the estimated two-level model. Unlike the omitted middle cluster case in Chapter 2, the omitted between-cluster variance repartition here is not weighted by the cluster size.

K (District)	J (School)	S (Student)	1	1	1	1	2	2	2	2	3	3	3	3
			1	1	2	2	3	3	4	4	5	5	6	6
			1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	ρ_1	ρ_2	ρ_2	0	0	0	0	0	0	0	0
1	1	2	ρ_1	1	ρ_2	ρ_2	0	0	0	0	0	0	0	0
1	2	3	ρ_2	ρ_2	1	ρ_1	0	0	0	0	0	0	0	0
1	2	4	ρ_2	ρ_2	ρ_1	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	ρ_1	ρ_2	ρ_2	0	0	0	0
2	3	6	0	0	0	0	ρ_1	1	ρ_2	ρ_2	0	0	0	0
2	4	7	0	0	0	0	ρ_2	ρ_2	1	ρ_1	0	0	0	0
2	4	8	0	0	0	0	ρ_2	ρ_2	ρ_1	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	ρ_1	ρ_2	ρ_2
3	5	10	0	0	0	0	0	0	0	0	ρ_1	1	ρ_2	ρ_2
3	6	11	0	0	0	0	0	0	0	0	ρ_2	ρ_2	1	ρ_1
3	6	12	0	0	0	0	0	0	0	0	ρ_2	ρ_2	ρ_1	1

K (District)	J (School)	S (Student)	1	1	1	1	2	2	2	2	3	3	3	3
			1	1	2	2	3	3	4	4	5	5	6	6
			1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	ρ	0	0	0	0	0	0	0	0	0	0
1	1	2	ρ	1	0	0	0	0	0	0	0	0	0	0
1	2	3	0	0	1	ρ	0	0	0	0	0	0	0	0
1	2	4	0	0	ρ	1	0	0	0	0	0	0	0	0
2	3	5	0	0	0	0	1	ρ	0	0	0	0	0	0
2	3	6	0	0	0	0	ρ	1	0	0	0	0	0	0
2	4	7	0	0	0	0	0	0	1	ρ	0	0	0	0
2	4	8	0	0	0	0	0	0	ρ	1	0	0	0	0
3	5	9	0	0	0	0	0	0	0	0	1	ρ	0	0
3	5	10	0	0	0	0	0	0	0	0	ρ	1	0	0
3	6	11	0	0	0	0	0	0	0	0	0	0	1	ρ
3	6	12	0	0	0	0	0	0	0	0	0	0	ρ	1

Figure 4.3 Correlation structures of $\boldsymbol{\psi}_K$ of the three-level model, and $\tilde{\boldsymbol{\psi}}_J$ of the two-level model omitting the highest cluster level.

4.3.2 Quantifying the Standard Error Estimate Bias

Bias of the Standard Error Estimates of the Coefficients of $Z_{j(k)}$ and $Z_{i(jk)}$

Predictor $Z_{j(k)}$, though a measure of the districts, is falsely disaggregated at the school level. The letters in the parentheses (i.e., (k) and (jk)) indicate the corresponding omitted cluster levels.

The estimated variance of the coefficient parameter of $Z_{j(k)}$ in the two-level model is:

$$Var_{2L}(\gamma_{02}) = \left\{\sum_{j=1}^{M_J}({Z_{j(k)}}' \tilde{\boldsymbol{\psi}}_{\boldsymbol{K}} Z_{j(k)})\right\}^{-1} = \sigma^2 \tilde{\tau}_{Z_{j(k)}} \left\{\sum_{j=1}^{M_J}({Z_{j(k)}}' Z_{j(k)})\right\}^{-1},$$

where $\tilde{\tau}_k = 1 + (n_L - 1)\rho$. In the three-level model which correctly models the predictor $Z_{j(k)}$ as Z_k, the variance estimate of the coefficient parameter is

$$Var_{3L}(\gamma_{001}) = \left\{\sum_{k=1}^{K}(Z_K{}'\boldsymbol{\psi_K}\, Z_K)\right\}^{-1} = \sigma^2\tau_{Z_k}\left\{\sum_{k=1}^{K}(Z_k{}'\, Z_k)\right\}^{-1},$$

where $\tau_{Z_k} = 1 + (n_L - 1)\rho_1 + n_L(n_H - 1)\rho_2$. The inflation of $Var_{2L}(\gamma_{02})$ is then quantified by the division of $Var_{3L}(\gamma_{001})$ and $Var_{2L}(\gamma_{02})$, which yields to

$$VOC_H^{(3-2,2L)} = \frac{Var_{3L}(\gamma_{001})}{Var_{2L}(\gamma_{02})} = \frac{\tau_{Z_k}}{\tilde{\tau}_{Z_{j(k)}}} = \frac{1 + (n_L - 1)\rho_1 + n_L(n_H - 1)\rho_2}{1 + (n_L - 1)\rho}$$

$$= 1 + \frac{n_L(n_H - 1)\rho_2}{1 + (n_L - 1)\rho}. \quad (4.5)$$

Noticeably, $VOC_H^{(3-1)}$ is identical to $VOC_M^{(3-1)}$ in Chapter 2 since both of them solve the same issue of adjusting the standard error estimates of the highest-level predictor coefficients when the two layers of clustering are omitted.

Bias of the Standard Error Estimates of the Coefficients of $W_{j(k)}$ and $W_{i(jk)}$

The school-level predictor $W_{j(k)}$ is the main predictor of interest[13]. Its coefficient's variance estimate inflation is quantified via comparing the variance estimate from a satisfactory two-level model where no higher cluster level exists and a false two-level model where a higher level exists but is omitted. The satisfactory two-level model is identical to the case that has been illustrated in Chapter 2 in deriving for $VOC_M^{(2-1,OLS)}$. The corresponding variance estimate of W_j's coefficient γ_{01} is

[13] The following derivation process applies to both cases of $W_{j(k)}$ as a continuous measure or binary treatment assignment indicator. The latter applies to the previous theoretical discussion of when the estimated model is a two-level random intercept CRT omitting the highest cluster level and the true model is a three-level random intercept BRT. When the true three-level BRT model has no random slope of $W_{j(k)}$, the standard error estimate of the difference between means is the same as the one from the three-level random intercept CRT model (see Konstantopoulos, 2008a and Konstantopoulos, 2008b). The equivalence of the standard error estimates of the continuous and binary predictors' coefficients has been shown in Section 2.3, Chapter 2.

$$\widetilde{Var}_{2L}(\gamma_{01}) = \left\{\sum_{j=1}^{J}(W_j{}'\,\boldsymbol{\psi_J} W_j)\right\}^{-1} = \sigma^2 \check{\tau}_{W_j}\left\{\sum_{j=1}^{J}(W_j{}'\,W_j)\right\}^{-1},$$

where $\boldsymbol{\psi_J}$ is the error variance-covariance matrix

$$\boldsymbol{\psi_J} = \sigma^2\begin{bmatrix} 1 & \rho_0 & \cdots & \rho_0 \\ \rho_0 & 1 & \cdots & \rho_0 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_0 & \rho_0 & \cdots & 1 \end{bmatrix} = \sigma^2\left[(1-\rho_0)\boldsymbol{I} + \rho_0\, \boldsymbol{l}_{n_L}\boldsymbol{l}'_{n_L}\right].$$

The corresponding variance weighting index is $\check{\tau}_{W_j} = 1 + (n_L - 1)\rho_0$, where the only intraclass correlation is $\rho_0 = \frac{\sigma^{(j)}}{\sigma^2}$, when the district-level does truly not exist. In the false two-level model, the error variance-covariance matrix is $\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}}$, which gives the variance estimate of $W_{j(k)}$'s coefficient γ_{01}:

$$Var_{2L}(\gamma_{01}) = \left\{\sum_{j=1}^{J}(W_{j(k)}{}'\,\widetilde{\boldsymbol{\psi}}_{\boldsymbol{J}} W_{j(k)})\right\}^{-1} = \sigma^2 \tilde{\tau}_{W_{j(k)}}\left\{\sum_{j=1}^{J}(W_{j(k)}{}'\,W_{j(k)})\right\}^{-1},$$

where $\tilde{\tau}_J = 1 + (n_L - 1)\rho$. Finally, the variance inflation factor yields to

$$VIF_H^{(2-2,2L)} = \frac{\widetilde{Var}_{2L}(\gamma_{01})}{Var_{2L}(\gamma_{01})} = \frac{\check{\tau}_{W_j}}{\tilde{\tau}_{W_{j(k)}}} = \frac{1 + (n_L - 1)\rho_0}{1 + (n_L - 1)\rho} = 1 - \frac{(n_L - 1)\rho_2}{1 + (n_L - 1)\rho}. \qquad (4.6)$$

As shown, when the district level cluster should be modeled as random effects but is omitted, the standard error estimates of $W_{j(k)}$'s fixed effect is overestimated and lead to Type II error. This finding is similar to $VIF_M^{(1-1,2L)}$ that was developed for the individual-level predictor $X_{i(j)k}$ in Chapter 2. The common idea is that, if the satisfactory model is a two-level, then the artificial between-group variance of the untrue highest level should be taken out.

Further, when both sources and layers of the clustering dependency are completely omitted in a single-level analysis using OLS estimation and W_{jk} is disaggregated at the student level as $W_{i(jk)}$, the corresponding variance inflation factor is

$$VOC_H^{(2-1,OLS)} = \tau_j = 1 + (n_L - 1)\rho_0, \tag{4.7}$$

which is identical to $VOC_M^{(2-1,OLS)}$ in Chapter 2 and is smaller $VOC_H^{(3-1,OLS)}$ by $(n_0 - 1)\rho_2$.

Bias of the Standard Error Estimates of the Coefficients of $X_{ij(k)}$ and $X_{i(jk)}$

Finally, since the individual-level variance is not affected by the omitted highest cluster level, the standard error estimate of $X_{ij(k)}$'s coefficient γ_{10} remains unbiased. This can be shown by

$$VOC_H^{(1-1,2L)} = \frac{1 - \rho_0 - \rho_2}{1 - \rho_1} = \frac{1 - \rho_1}{1 - \rho_1} = 1. \tag{4.8}$$

In terms of OLS estimation, the $VOC_H^{(1-1,OLS)}$ is identical to the $VOC_M^{(2-1,OLS)}$ in Chapter 2 (see Eq. 2.9) that

$$VOC_H^{(1-1,OLS)} = 1 - \rho_0 - \rho_2. \tag{4.9}$$

Table 4.1 below summarizes the above derived variance inflation factors corresponding to the predictors of each level in the estimated two-level HLM and single-level OLS models.

Table 4.1 A summary of VOCs when the highest cluster level is omitted in a three-level structured clustering data.

Three-level HLM		**Two-level HLM**			**Single-level OLS Estimation**		
Level	*Predictor*	*Level*	*Predictor*	*Variance adjustment*	*Level*	*Predictor*	*Variance adjustment*
Student	X_{ijk}	Student	$X_{ij(k)}$	$VOC_H^{(1-1,2L)} = 1$	Student	$X_{i(jk)}$	$VOC_H^{(1-1,OLS)} < 1$
School	W_{jk}	School	$W_{j(k)}$	$VOC_H^{(2-2,2L)} < 1$		$W_{i(jk)}$	$VOC_H^{(2-1,OLS)} > 1$
Districts	Z_k		$Z_{j(k)}$	$VOC_H^{(3-2,2L)} > 1$		$Z_{i(jk)}$	$VOC_H^{(3-1,OLS)} > 1$

Note. The letters in the parentheses indicate the corresponding cluster levels that are omitted.

4.3.3 Simulation Results

Similar to Chapter 2, a simulation study is designed to test the variance estimation bias when the highest cluster level is omitted as well as the performance of the derived VOC formulas. The total sample size of students (M_I) and number of schools (M_J) are fixed to be 2000 and 100, which lead to a conventional school size $n_L = 20$. Four conditions of number of districts (M_k) are set to be 5, 10, 25, and 50, which covers a plausible range of sample size of the highest cluster level.

In each condition of M_k, the residual variance ($\tilde{\sigma}^{(i)}$) and school-level random effect variance ($\tilde{\sigma}^{(j)}$) of the estimated two-level models are set with two pairs: 0.5 and 0.5, and 0.8 and 0.2. The latter pair setting satisfies the empirical evidence where between-school variance could reach to 0.2 (Hedges & Hedberg, 2014; Konstantopoulos, 2009; Westine et al., 2013). In Fahle and Reardon (2018), the between-districts variance $\sigma^{(k)}$ of U.S. public school for Grades 3-8 students in Math and English Language Arts ranges from 0.05 to 0.24. Then, the setting of the

omitted district-level random effect variance ($\sigma^{(k)}$) in this current study includes the evidence found in Fahle and Reardon (2018) and a hypothetical extreme large value, which are 0.1, 0.4, and 0.6. Setting $\sigma^2_{Total} = 1$, the values of random effects variance $\tilde{\sigma}^{(j)}$, $\sigma^{(k)}$, and $\sigma^{(j)}$ are equivalent to the ICCs of ρ (or ρ_1), ρ_2, and ρ_0, respectively. Again, the magnitude of the estimation bias and the performance of VOCs' adjustment of estimates are measured by the index of relative bias R. B.$_{est}$ and R. B.$_{adj.est}$ correspondingly. See Appendix 4.A for the parameter settings and simulation results.

Bias of the Random Effects and the Adjustment Performance

Previous research had found that the omitted $\sigma^{(k)}$ is taken by $\tilde{\sigma}^{(j)}$, while $\tilde{\sigma}^{(i)}$ remains the same. The simulation results support this finding. In all conditions, the mean R. B.$_{est}$ of $\tilde{\sigma}^{(i)}$ are all zero. With larger setting of $\sigma^{(k)}$ (or ρ_2) and M_k, the overestimated $\tilde{\sigma}^{(j)}$ has larger positive R. B.$_{est}$. In the extreme condition of $\sigma^{(k)} = \rho_2 = 0.6$, $\tilde{\sigma}^{(j)}$ can be more than twice and even four times larger than $\sigma^{(j)}$ as the number of schools increases. Even in the cases where $\sigma^{(k)}$ is extremely small of being 0.1, between-school variation can be overestimated by at least around 15%. With adjustment, the mean R. B.$_{adj.est}$ of $\tilde{\sigma}^{(j)}$ are close to 0 across all conditions.

Bias of the Standard Error Estimates of the Coefficients of $W_{j(k)}$ and $W_{i(jk)}$ and the Adjustment Performance

When the district-level cluster is omitted in the estimated two-level model, the positive values of mean R. B.$_{est}$ indicates that the standard error estimates of $W_{j(k)}$ is overestimated which lead to Type II error. The magnitude of the overestimation increases with the increase of $\sigma^{(k)}$ and M_k. When $\sigma^{(k)}$ is considerably large as 0.6, the standard error estimates of the two-level

model are 1.5 to 2 times larger than the parameter. When the omitted $\sigma^{(k)}$ is trivial as 0.1, the magnitude of the overestimation is minimal.

When both school- and district-level clusters are not modeled as in the single-level model, the standard error estimates of $W_{j(k)}$ are underestimated as the mean R. B.$_{est}$ are all negative. The value of R. B.$_{est}$ are relatively stable, around -0.6 across all conditions. This is because the setting of the overall omitted clustering dependency $\sigma^{(j)}$ are relatively similar of being 0.5 and 0.8, and the sample size of students is fixed.

Finally, for the adjustment performance, both $VOC_H^{(2)}$ and $VOC_H^{(2-1)}$ for the two- model and single-level model are desirable since the mean R. B.$_{adj.est}$ are consistently smaller than 0.1.

Bias of the Standard Error Estimates of the Coefficients of $Z_{j(k)}$ and $Z_{ij(k)}$ and the Adjustment Performance

When the district-level predictor Z_k is falsely disaggregated, either at the school-level in the two-level model or at the student- in the single-level model, the standard error estimates of the coefficient of Z_k are underestimated. In the two-level model, the underestimation bias increases with the increase of $\sigma^{(k)}$ and the decrease of M_k. With the maximal $\sigma^{(k)}$or $\rho_2 = 0.6$, the standard error estimates can be around 60% larger than the parameter. With the OLS estimation, the underestimation magnitude is relatively stable as the mean R. B.$_{est}$ are around -0.8 across all conditions. Again, this is due to the omission of the overall clustering dependency in the single-level analysis, regardless of the proportion of each cluster level's variance. Further, the underestimation magnitude in the OLS estimation is always higher than the two-level model in each condition. This is because $\tilde{\sigma}^{(j)}$ in the two-level models have captured the omitted $\sigma^{(k)}$.

The performance of the VOCs is ideal across all settings and models, where the mean R. B.$_{est}$ are close to or smaller than 0.1.

Bias of the Standard Error Estimates of the Coefficients of $X_{ij(k)}$ and $X_{i(jk)}$ and the Adjustment Performance

As shown, the standard error estimates of the coefficients of $X_{ij(k)}$ is not biased in the two-level model. However, the standard error estimates of the coefficients of $X_{i(jk)}$ it is overestimated using OLS estimation. This pattern is consistent with the previous findings in Chapter 2 that omitting the adjacent higher cluster level leads to Type II error issue. The overestimation is large when ρ_2 and M_k are large. For example, when the omitted $\sigma^{(k)}$ or ρ_2 is 0.6 and M_k is 50, the standard error estimates of the coefficients of $X_{i(jk)}$ are two times larger than the parameters. When the omitted $\sigma^{(k)}$ or ρ_2 is 0.1, the estimates can still be 40% larger than the parameters. $VOC_H^{(1-1,\mathrm{OLS})}$ performed relatively well that in most of the cases, the mean R. B.$_{est}$ are around or smaller than 0.1, though in two cases of when $M_k = 5$ and $\sigma^{(k)}$ are large (i.e., 0.6 and 0.4), the mean R. B.$_{est}$ are around 0.2.

4.4. Empirical Example and Sensitivity Analysis

This section employs the same study of Heafner et al. (2019) seen in Chapter 3 for example. As shown earlier, their employed data NAEP-E has a two-stage sampling design where schools are PSUs and students are SSUs, and the empirical model is a two-level random intercept HLM model. With a large sample size of schools $M_J = 56$, an incidental highest cluster level of districts could emerge. Further, as stated by the authors, state and district level policy predictors, including required economics education for graduation and economics testing, are modeled at the

school level (see Heafner et al., 2019, p. 334). In this case, these variables are falsely aggregated at the school level and produced with underestimated standard errors, though no significant evidence were found. Since Chapter 3 has already demonstrated examples of making Type I error, this section provides example of Type II errors of the middle-level predictors when a higher cluster level is omitted. The example predictor used here is the requirement of economics education for graduation, which I consider as a true school-level predictor for example-making reason.

The procedure of conducting the sensitivity analysis follows the steps shown in heuristics diagram of Figure 3.3, and the results are shown in Table 4.2. The regression coefficient and random effects estimates from the estimated two-level model are presented at the left upper corner in the table, where $t_{2L} < t^{\#}$ satisfies the requirements of conducting the Type II error sensitivity analysis. Since the omitted between-district variance ($\sigma^{(k)}$) is completely captured by the estimated between-school variance of the two-level model ($\tilde{\sigma}^{(j)}$) with no weights of cluster sizes, the sensitivity analysis here is straightforward and starts with calculating the threshold $\sqrt{VOC_t} = 0.432$ and the corresponding $\rho_2 = 0.267$.

For any settings of ρ_2 that is larger than 0.267 and $\sqrt{VIF_0}$ that is smaller than 0.432, the risk of having Type II error is larger than 0. For example, the maximum ρ_2 found is 0.275 with a corresponding $\sqrt{VOC_{max}} = 0.401$. That is, when the hypothesized district-level ICC ρ_2 is 0.275 or the estimated between-school variance are completely between-district variance, the magnitude of inference robustness (or effect size) reduces by 60% and the risk of making Type II error increases by 12% when compared with the threshold setting. In the current example, the maximum ρ_2 does not exceeded ρ of the estimated two-level model.

Table 4.2 Sensitivity analysis of the school-level predictor: economics required for graduation

Estimated Two-level HLM						**Hypothesized Three-level HLM**					
				M_J	560						
				n_L	20						
$\hat{\beta}$	1.820			$\tilde{\sigma}^{(i)}$	22.62	$\sigma^{(i)}$	22.620	22.620	22.620	22.620	22.620
$StE^{\#}$	0.929	$t^{\#}$	1.96	$\tilde{\sigma}^{(j)}$	8.58	$\sigma^{(j)}$	0.265	2.340	8.268	0.000	-0.780
StE_{2L}	2.150	t_{2L}	0.85	/	/	$\sigma^{(k)}$	8.315	6.240	0.312	8.580	9.360
$\hat{\beta}$	1.820			ρ	0.275	ρ_0	0.008	0.075	0.265	0.000	-0.025
						ρ_2	0.267	0.200	0.010	0.275	0.300

$\boldsymbol{\rho_2}$		$\sqrt{\boldsymbol{VOC}}$		$\boldsymbol{StE_{voc}}$	$\boldsymbol{W_{OC}}$ & $\boldsymbol{ES_{OC}}$	$\boldsymbol{R_{OC}}$
		Threshold $\sqrt{VOC_t}$	0.432	0.929	0.568	*Switch Point*
$\sqrt{VOC_t}$ based ρ_2	0.267	$\sqrt{VOC_0}$	0.432	0.929	0.568	*Switch Point*
ρ_2	0.200	$\sqrt{VOC_0}$	0.624	1.342	0.376	*NA*
$\rho_{2,min}$	0.010	$\sqrt{VOC_{min}}$	0.985	2.117	0.015	*NA*
$\rho_{2,max}$	0.275	$\sqrt{VOC_{max}}$	0.401	0.862	0.599	0.121
Not plausible	0.300	*NA*	0.290	0.624	0.710	0.462

An implausible example of $\rho_2 = 0.3$ is thus demonstrated that if $\rho_2 > \rho$, the between-school variance from the three-level models turns to be negative, though the corresponding robustness measures are producible and larger than the above ones. Also, $\rho_{2,min}$ is provided to show the lower boundary of the variance adjustment. In this case, the reduced robustness and effect size is small (i.e., 1.5%) , thus no concerns for making Type II error. The above hypothesized ρ_2 are in a comparable range of around 0.05 to 0.24 and are of the empirical values summarized in previous literature across nations, grade level, and subjects (e.g., Fahle & Reardon, 2017). This evidence heightens the significance of conducting this sensitivity analysis to test the estimation bias due to an omitted but empirically possible district cluster level.

4.5 Discussion and Conclusion

This chapter clarifies the risky practice in two-level models omitting the highest cluster level that is legitimate in sampling and experimental designs. The harmful ramifications of omitting the clustering dependency of the highest level need particular caution from researchers when their research questions relate to the cluster-level predictors effect, and to explaining the clusters' capability in demonstrating the error variance of the individual variance, since the estimated two-level model would generate biased standard error estimates of the middle and highest level coefficients estimates and random effect variance of the middle cluster level. Similar to Chapter 2, the VOCs derived in this chapter quantify the potential magnitude of the estimation bias of each cluster levels' predictors that can be applied to general modeling settings.

The decision on whether to explicitly model the highest cluster level depends on the research design of sampling and experimental schemes, as well as the rationales of whether to generalize the reference to the studied sample groups only or to the population of interest. When the main predictor of interest is at the middle level and the highest level of clusters are the population groups, a fixed effect modeling framework is genuine. In contrast, if the predictors of interest also include the highest-level ones and the clusters are sample units, the highest cluster level needs to be modeled as a random effect. As listed in Table 4.1, the estimated two-level model omitting the highest cluster level would lead to a Type I error of the middle-level predictor estimate and a Type II error of the disaggregated highest-level predictor. However, the individual level inferences would not be affected. The extended single-level model scenario of omitting the overall clustering dependency has been shown in Chapter 2 where Type I error emerges for the cluster-level predictors estimates and Type II errors for the individual level predictor.

Another particularity of making decisions on modeling the highest cluster level relates to the sample size. Frequently, the small sample size issue happens at the highest cluster levels, particularly when the highest cluster level is not in the initial sampling design. In this case, the fixed-effect approach is optimal (McNeish & Wentzel, 2016; McNeish & Kelley, 2019). The current study only tested when the middle level cluster size is relatively large (i.e., fixed the school size n_L to be 20) and the cluster size of the highest cluster level is not extremely small (where the smallest district sample size M_K is 5), the displayed simulation outputs did not evidence any exceptional performance of the VOCs relating to the sample size. In future studies, the small sample size occasions relevant to the assumption of random effects and performance of the estimation methods should be investigated in detail, whereas it is out of scope of the current study. Also, future studies are encouraged to develop extended VOCs in conditions of unbalanced design and random slopes.

CHAPTER 5

OMITTED SERIAL CORRELATIONS IN LOWEST CLUSTER LEVEL

5.1 Introduction

Longitudinal data can be conceptualized as clustered data, since repeated measures are clustered within groups such as the yearly measured performance of students. A two-level linear growth model is commonly utilized to describe students' average performance change and the change variabilities, as well as examine the factors that can explain the growth patterns (Bryk & Raudenbush, 2002; Hoffman, 2015; Singer & Willett, 2003).

In previous chapters, units within groups are exchangeable in conventional clustered data that any pair of units within a cluster has an equal intraclass correlation as they are assumed to share common unobserved factors at the group level (Alejo et al., 2018; Cameron & Miller, 2015; Hansen, 2007). Assuming homogeneity and independent two levels of random effects, the corresponding error variance-covariance of a two-level model is $\boldsymbol{\psi}_i = \text{var}(Y_i) = \mathbf{R} + \mathbf{l}_{n_c}\mathbf{G}\mathbf{l}'_{n_c}$, where $\mathbf{R} = \sigma^{(i)}\mathbf{I}$ is the first-level error structure. The second-level error structure $\mathbf{G}$ is a $\mathbf{l}_{n_c} \times \mathbf{l}'_{n_c}$ matrix of $\sigma^{(j)}$, where $\mathbf{l}_{n_c}$ is a column vector of one with a length of cluster size n_c,

A distinguished feature of longitudinal data is that repeated measures are chronologically ordered (Alejo, et al., 2018; Skrondal & Rabe-Hesketh, 2008). The ordering gives an additional source of dependency from the correlations of repeated measures within individuals of an outcome, other than the mean differences across individuals and the variations of growth across individuals (Hoffman, 2015). Unlike equicorrelated intraclass correlations, serial correlations between two successive time measures (i.e., $\text{corr}(y_{(t-1)i}, y_{ti})$) have certain

patterns. Generally, the correlations between two successive time measures are larger than the correlations between two non-successive ones[14]. As the gap between two occasion measures increases, the correlation decreases. That is, $\mathrm{corr}(y_{(t-k)i}, y_{ti}) > \mathrm{corr}(y_{(t-k)i}, y_{(t+s)i})$. In this case, another form of intraclass correlation due to serial correlation emerges, in addition to the conventional one due to random effects. The basic identity structure (ID) of $\mathbf{R} = \sigma^{(i)}\mathbf{I}$, assuming independently and identically distributed within-individual-repeated-measure residuals, is then overly simplified in the multilevel longitudinal data analysis.

In the field of Economics, the intraclass correlations due to clustering and serial correlation are explicitly defined to be closely related but distinct (Angrist & Pischke, 2008). Then corresponding statistical tests are proposed for evaluating the two forms of intraclass correlations in random effects longitudinal models. As surveyed in Alejo et al. (2018), earlier tests evaluating either random effects or serial correlation (i.e., Baltagi & Li, 1991; Breusch & Pagan, 1980) tend to produce inflated rejection rate if the other form of intraclass correlation exits and is ignored (Bera et al., 2001). Empirical research also presents this issue. In the influential study of Bertrand et al. (2004), a survey of 92 difference-in-difference (DD) studies found that only five of them had implemented serial correlation corrections. In that study, significant over-rejection is found for a null effect treatment, which is due to the omitted serial correlation. On the other hand, interclass correlation due to clustering alone is commonly taken care of by cluster-robust standard errors (Moulton, 1986, 1990), alternative estimators such as GLS (Liang & Zeger, 1986; White, 1980), and block bootstrap (Cameron et al., 2008). Later developed tests provide joint tests of both forms of intraclass correlation such as in Alejo et al.

[14] The current study focuses on positive serial correlations, which means the error terms have the same sign from one time-measure to the next.

(2018)**,** Baltagi, Jung, and Song, (2002, 2010), and King and Roberts (2015), to name a few**.** These studies highlight the identification of the sources of the intraclass correlation (i.e., due to clustering effect or serial correlation), and appropriate strategies and models would be applied (Alejo et al., 2018). With both forms of intraclass correlations, corresponding strategies, such as feasible generalized least squares estimation (FGLS) (Hansen, 2007), are required to account for dependencies (Angrist & Pischke, 2008; Wooldridge, 2003).

The above discussion alerts the critical need for detecting and distinguishing the two forms of intraclass correlation. Beyond the above-mentioned approaches that are popular in economics research, the model-based approach HLM account for the two forms of intraclass correlation simultaneously by specifying a correct error variance-covariance structure. However, it is not uncommon in empirical research that the serial correlations among the repeated measures are ignored in the time-level variance $\boldsymbol{R}$, and all the expected correlations among the repeated measures are (false) due to the individual-level random effect variance $\boldsymbol{G}$**.** Consequently, the tested theories and inferences made for the variance components and fixed effects could be erroneous (Ferron et al., 2002; Hoffman, 2015; LeBeau, 2016, 2018). Therefore, with recognition of serial correlation, a correctly specified $\boldsymbol{R}$ structure is pivotal.

As a start, the current study considers the ID structure of $\boldsymbol{R}$ being a scenario of omitting serial correlations at the lowest level, and sets out to mathematically quantify the corresponding estimation bias for robust inference making. It begins in Section 5.1 with a review of the approaches to specify $\boldsymbol{R}$, and a discussion of the bias in estimates due to the misspecified $\boldsymbol{R}$ in empirical research. Then this article's study motivation and goal is proposed. Section 5.2 follows the details of deriving generalized formulas to quantify the estimation bias of variances of the random effects and fixed effects, explore through an example of a two-level random intercept

linear growth model that misspecified the $\boldsymbol{R}$ as ID from AR (1). A Monte Carlo simulation study presents the performances of the formulas. Section 5.4 provides an empirical study example. At last, Section 5.5 concludes and discusses the future research.

5.2 Alternative $\boldsymbol{R}$ Structures with Serial Correlations

The structure of the time-level residual variance matrix $\boldsymbol{R}$ represents the serial correlation patterns. Besides the ID structure of $\boldsymbol{R} = \sigma^{(i)}\boldsymbol{I}$, many alternative structures have been widely recognized in textbooks of multilevel analysis, including Bryk & Raudenbush (2002, Chapter 6), Hoffman (2015, Section 3), Singer & Willett (2003, Chapter 7), and Snijders & Bosker (2012, Chapter 15), to name a few. Commonly presented alternative $\boldsymbol{R}$ structures include autoregression (AR (k)), autoregression and moving average (ARMA (p, q)), Toeplitz (TOEP(k)), and unstructured.

In practice, the selection of $\boldsymbol{R}$ largely depends on empirical and theoretical needs (Snijders & Bosker, 2012). Nevertheless, this approach is limited by prior experience and generalizability, which is prone to uncertainties in specifying $\boldsymbol{R}$. Moreover, a misspecified $\boldsymbol{R}$, in return, distorts the deduction of theories. Taking a simple example, which will be proved in later sections, when a relatively large serial correlation is completely omitted, the between-individual variance matrix $\boldsymbol{G}$ is then considerably overestimated as $\boldsymbol{R}$ is underestimated. Then, in modeling, individual-level predictors are added to explain the overstated between-individual variances, instead of the within-individual predictors (Hoffman, 2015). In this case, the true predictors and mechanisms of individual growth, particularly for the within-individual levels, are overlooked. This example applies well for research that is interested in examining the impacts of students' time-varying psychological and cognitive factors on their learning. On the other hand, if the

serial correlation is overstated and the research is interested in students' attributes, like ethnicity and family background, an overstated $\boldsymbol{R}$ eliminates some between-individual variance due to those students' attributes.

In addition to deciding $\boldsymbol{R}$ based on empirical experience and theory, a general statistical approach is through comparing the goodness of fit values among several models with different specified $\boldsymbol{R}$ structures, and then selecting the best fit model. However, the arbitrary values of likelihood ratio tests and information fit criteria have been critiqued. The criteria of modeling performances depend on many factors, including the number of time measures, total sample size of individuals, estimation methods, and variance-covariance patterns. Therefore, no single criterion performs uniformly better than the others, and certain criteria perform better for selecting some $\boldsymbol{R}$ structure models (Vallejo et al., 2011). Also, it is important to note that the best fit model is not necessarily the model with the correct $\boldsymbol{R}$ (Murphy & Pituch, 2009). Researchers may turn to the general unstructured $\boldsymbol{R}$ with no prior specifications to best fit the data (Littell et al., 2000). However, the unstructured $\boldsymbol{R}$ is less interpretable to empirical studies that appreciate substantive theories. Further, as evidenced in Murphy and Pituch (2009), although the unstructured $\boldsymbol{R}$ produces the least biased random effects, it inflates Type I error rate of fixed effects and has convergence problems as a large number of parameters needs to be estimated.

The above presented selection methods of $\boldsymbol{R}$ are not free from concerns. Empirical research is then still subject to the serious impacts on variance estimates if $\boldsymbol{R}$ is misspecified. In Kwok et al. (2007) study, three scenarios of misspecifying $\boldsymbol{R}$ are summarized: overspecification, underspecification, and general misspecification. That study develops a network of multiple $\boldsymbol{R}$ by their nesting relationship of structures, including independent (ID), first-order autoregressive process (AR(1)), first-order moving average process ARMA (1,1), Toeplitz 2 bands (TOEP(2)),

and unstructured, as shown in Figure 1 and Table 1 (Kwok et al., 2007, p.565, 568). For example, an underspecification situation happens when the true $\boldsymbol{R}$ is AR(1) while an ID is estimated, or the true one is ARMA(1,1) while an AR(1) is estimated. An overspecification happens when AR(1) is the true $\boldsymbol{R}$ while ARMA(1,1) is modeled. TOEP(2), and unstructured are considered as general misspecification when the true $\boldsymbol{R}$ is defined as the other ones. The general findings are that, if $\boldsymbol{R}$ is underspecified or general-misspecified, the fixed effects coefficients' estimates are unbiased, while the variances are found to be overestimated. The overspecifications lead to slightly underestimated variances.

However, other studies found conflicting patterns. Murphy and Pituch (2009) detects smaller standard error estimates in the underspecified AR (1) model while the true $\boldsymbol{R}$ is ARMA (1, 1). Also, Ferron et al. (2002) finds larger estimates of random effects' variances from the estimated ID model when the true $\boldsymbol{R}$ is AR (1), whereas the standard error estimates of the fixed effects are slightly smaller than they should be, as the Type I error rate inflates accordingly. These finding are consistent with a recent Monte Carlo study of LeBeau (2018), which also shows inflated Type I error rates of the fixed effects when the serial correlation is completely omitted in $\boldsymbol{R}$ (i.e., underspecified as ID).

The above simulation-based studies provide evidence of estimation bias due to the misspecified $\boldsymbol{R}$, whereas the findings are not always consistent. Moreover, they are limited in generalizability as they are tested for certain range of parameters. Therefore, further analytic examinations are needed to further determine the underlying mechanisms of the estimation bias.

5.2.1 Study Motivation

The above discussion demonstrates that the decision making of $\boldsymbol{R}$ structure is complex. Besides the awareness of the negative impacts of misspecified $\boldsymbol{R}$**,** empirical researchers can benefit more from a strategy that aids to evaluate whether the estimated $\boldsymbol{R}$ is specified correctly, and to adjust the potential bias if the estimated $\boldsymbol{R}$ is false.

The current study intents to provide such a strategy that, instead of deciding *the* true $\boldsymbol{R}$, it proposes to quantify the uncertainties that the specified $\boldsymbol{R}$ in the estimated model can cause when an alternative $\boldsymbol{R}$ is hypothesized to be true. Bertrand et al. (2002) provides variance estimate formulas to demonstrate the exact reason of omitting positive serial correlation in OLS estimation that understates the standard error estimates. However, no such efforts have been made with the presentation of clustering dependencies in multilevel longitudinal analysis. The current study therefore contributes to fill this gap by deriving formulas to determine the reason of estimation bias due to omitted serial correlation with multilevel longitudinal analysis. The detected bias, then, can be adjusted by those formulas. These formulas will be derived similarly with the VOCs from the previous chapters of the omitted middle and highest cluster levels. This quantification approach distinguishes the sources of the estimation bias (i.e., serial correlation and random effects' variances), and the varying impacts of misspecified $\boldsymbol{R}$ on different levels of predictors, including the growth parameters and the time-varying predictors at the time-level, and the time-invariant predictors at the individual-level. In this way, researchers in practice can benefit from model building with selecting predictors that best explain those corresponding variances, as well as deciding whether a predictor's standard error estimates need adjustment or not.

Together with the sensitivity analysis developed in Chapter 3, this approach provides researchers with flexibilities to choose the best model that is statistically robust and appropriate for their theories. This approach also facilitates researchers and readers who do not have the original data to replicate the presented models while the original study does not provide much information on the selection criteria and decision rationale of $\boldsymbol{R}$. For instance, the assumption of $\boldsymbol{R}$ is not explicitly given in the study of a five-year longitudinal study of student achievement and goal setting (i.e., Moeller et al., 2012), and the model results do not show serial correlation estimates. In this case, readers may suspect the estimated model is specified $\boldsymbol{R} = \sigma^{(i)}\boldsymbol{I}$, and ask questions of, if any serial correlation is omitted, how much the estimation bias would be and how robust the inferences that were made.

Since ID and AR(1) are the most widely used $\boldsymbol{R}$ in empirical longitudinal research, the current study focuses on this underspecification case of estimated $\boldsymbol{R}$ being ID while the true one being AR(1). However, the above described approach is suitable to test many other pairs of misspecification cases, such as between AR(1) and ARMA(1,1), as long as the structures are nested as shown in Figure 1 of Kowk, et al. (2007). The current study adapts this concept of nested $\boldsymbol{R}$ structures for future work of building a full network of $\boldsymbol{R}$ structure misspecification pairs.

5.3 Quantification of Standard Error Bias

5.3.1 Model Setting

This section derives formulas to quantify the bias of variance estimates of both random effects and fixed effects, if the true $\boldsymbol{R}$ structure is assumed to be AR (1) and the estimated $\boldsymbol{R}$ is

ID. The following defines a two-level random intercept linear growth model to describe a mean pattern of students' growth over time:

$$\text{Time-level: } Y_{ti} = \beta_{0i} + \beta_{1i} Time_{ti} + \varepsilon_{ti},$$

$$\text{Student-level: } \beta_{0i} = \gamma_{00} + \gamma_{01} X_i + r_{0i},$$

$$\beta_{1i} = \gamma_{10},$$

$$\text{Mixed model: } Y_{ti} = \gamma_{00} + \gamma_{10} Time_{ti} + \gamma_{01} X_i + r_{0i} + \varepsilon_{ti}.$$

For simplicity of notation and formula derivation, the model is a balanced design that any student i is measured at the same n_t occasions. The intercept β_{0i} varies at the student-level and is explained by a student-level measure X_i. The occasion measure is $Time_{ti}$ and its' coefficient parameter γ_{10} is the mean growth rate of students. Taking five-year measured age for example, $Time_{ti}$ can be coded as 1, 2, 3, 4, 5, or -2, -1, 0, 1, 2, where the 0 point serves a meaningful start point for interpretation (Hoffman, 2015). In here and later simulation, $Time_{ti}$ is group-centered which helps avoid endogeneity issues where random effects correlate with predictors (Antonakis et al., 2019). Though a random slope is common in longitudinal data analysis, the growth rate γ_{10} in this study is not assumed to be random, as students grow at a same rate in a shared context, such as the same school. Assuming the true serial correlation pattern is AR(1), the random effects are $\varepsilon_{ti} \sim N\left(0, \boldsymbol{R}_{\boldsymbol{AR(1)}}\right), r_{0i} \sim N\left(0, \sigma^{(i)}_{AR(1)}\right)$, and $cov(\varepsilon_{ti}, r_{0i}) = 0$.

Consistent with the modeling settings in Chapter 2 and 4, homogeneity assumption holds for both levels that random effects' variances are constant conditioning on controlled variables. In the model above, covariates other than $Time_{ti}$ and X_i are not shown for simplicity. In this

case, the fixed growth rate and serial correlation are also conditional. The model setting serves a presumption that the empirical model has no omitted confounding variable.

5.3.2 Quantifying the Standard Error Estimate Bias

The error variance-covariance structure is $\boldsymbol{\psi_{i,AR(1)}} = var(Y_i) = \boldsymbol{R_{AR(1)}} + \boldsymbol{l}_{n_t}\boldsymbol{G}\boldsymbol{l}'_{n_t}$, where the dimension of $\boldsymbol{\psi_{i,AR(1)}}$ is $n_t * n_t$, and $\boldsymbol{l}_{n_t}$ is a column vector of n_t ones. The difference of the error variance-covariance structure between the estimated ID model (noted as $\widetilde{\boldsymbol{\psi}}_{i,ID}$) and the AR (1) model is at the structure of $\boldsymbol{R_{AR(1)}}$. In the AR(1) model,

$$\boldsymbol{R_{AR(1)}} = \sigma^{(t)}_{AR(1)} \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}.$$

That is, with an AR(1) serial correlation pattern, the variance of time-level residual is $var(\varepsilon_{ti}) = \sigma^{(t)}_{AR(1)}$ and the covariance of two adjacent time measures is $cov(\varepsilon_{ti}, \varepsilon_{si}) = \sigma^{(t)}_{AR(1)}\rho^{|t-s|}$, where $t = 1,2,\ldots,n_t$; $s = t-1$, $\forall i, s \neq t$, and $\rho^{|t-s|} = \rho^{|s-t|}$ (Montes-Rojas, 2016). I also assume that no measurement error and the lag-1 autocorrelation is positive (i.e., $0 \leq \rho^{|t-s|} \leq 1$). The structure $\boldsymbol{l}_{n_t}\boldsymbol{G}\boldsymbol{l}'_{n_t}$ does not differ in the true AR(1) model or in the estimated ID model, which captures the intraclass correlation due to the individual-level random effect variance. The complete extended form of $\boldsymbol{\psi}_{i,AR(1)}$ is

$$\boldsymbol{\psi}_{i,AR(1)} = \boldsymbol{R_{AR(1)}}+\boldsymbol{l}_{n_t}\boldsymbol{G}\boldsymbol{l}_{n_t}{}'=\sigma^{(t)}_{AR(1)}\begin{bmatrix}1 & \rho^1 & \rho^2 & \rho^3\\ \rho^1 & 1 & \rho^1 & \rho^2\\ \rho^2 & \rho^1 & 1 & \rho^1\\ \rho^3 & \rho^2 & \rho^1 & 1\end{bmatrix}+\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\left[\sigma^{(i)}_{AR(1)}\right][1 \quad 1 \quad 1 \quad 1]$$

$$=\begin{bmatrix}\sigma^{(t)}_{AR(1)}+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^2+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^3+\sigma^{(i)}_{AR(1)}\\ \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^2+\sigma^{(i)}_{AR(1)}\\ \sigma^{(t)}_{AR(1)}\rho^2+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)}\\ \sigma^{(t)}_{AR(1)}\rho^3+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^2+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}\rho^1+\sigma^{(i)}_{AR(1)} & \sigma^{(t)}_{AR(1)}+\sigma^{(i)}_{AR(1)}\end{bmatrix}.$$

Unlike $\boldsymbol{\psi}_{i,ID}$, the off-diagonal of $\boldsymbol{\psi}_{i,AR(1)}$ is no longer a single $\sigma^{(i)}_{AR(1)}$ but a function of $\sigma^{(t)}_{AR(1)}$, $\rho^{|t-s|}$, and $\sigma^{(i)}_{AR(1)}$. To achieve a simpler form of $\boldsymbol{\psi}_{i,AR(1)}$ that can be written into a general linear form like $\boldsymbol{\psi}_{i,ID}$, and to achieve a general form of the column sum (such as τ^* in previous omitted middle and highest level cases), I construct an average term of $cov(\varepsilon_{ti}, \varepsilon_{si})$ in the off-diagonal as

$$\overline{cov}(\varepsilon_{ti}, \varepsilon_{si}) = \sigma^{(t)}_{AR(1)}\bar{\rho},$$

where

$$\bar{\rho} = \frac{\sum_{|t-s|=1}^{n_t-1}\rho^{|t-s|}}{n_t * n_t - n_t} = \frac{2}{n_t(n_t-1)}\boldsymbol{S}$$

and

$$\boldsymbol{S} = \sum_{t=\mathbf{2}}^{n_t}\sum_{s=t-1}^{n_t-1}\rho^{|t-s|}$$

is the sum of all elements of either side off-diagonal of the $\rho^{|t-s|}$ symmetric correlation matrix. This averaging approach is also suggested by Montes-Rojas (2016).

Any elements in the off diagonal of $\boldsymbol{\psi}_{i,AR(1)}$ now turns to $\sigma_{AR(1)}^{(t)}\bar{\rho} + \sigma_{AR(1)}^{(i)}$. Similar to the construction of the conventional ICC as shown in Appendix 2.1, the expected intraclass correlations of $\rho_{AR(1)}$ is defined by the ratio of $\sigma_{AR(1)}^{(t)}\bar{\rho} + \sigma_{AR(1)}^{(i)}$ and the total error variance

$$\rho_{AR(1)} = \frac{cov\left(Y_{ti} - \hat{Y}_{ti}, Y_{si} - \hat{Y}_{si}\right)}{\sqrt{var\left(Y_{ti} - \hat{Y}_{ti}\right)var\left(Y_{si} - \hat{Y}_{si}\right)}} = \frac{cov(\varepsilon_{ti} + r_{0i}, \varepsilon_{si} + r_{0i})}{\sqrt{var(\varepsilon_{ti} + r_{0i})var(\varepsilon_{si} + r_{0i})}}$$

$$= \frac{cov(\varepsilon_{ti}, \varepsilon_{si}) + cov(r_{0i}, \varepsilon_{si}) + cov(r_{0i}, \varepsilon_{ti}) + var(r_{0i})}{\sqrt{var(\varepsilon_{ti} + r_{0i})var(\varepsilon_{si} + r_{0i})}}$$

$$= \frac{\sigma_{AR(1)}^{(t)}\bar{\rho} + \sigma_{AR(1)}^{(i)}}{\sigma_{AR(1)}^{(t)} + \sigma_{AR(1)}^{(i)}} = \frac{\sigma_{AR(1)}^{(t)}\bar{\rho} + \sigma_{AR(1)}^{(i)}}{\sigma_{total}^{2}}$$

$$= \left(1 - \rho_{1,AR(1)}\right)\bar{\rho} + \rho_{1,AR(1)}. \quad (5.1)$$

Straightforwardly, $\rho_{AR(1)}$ is a function of the two forms of intraclass correlations from the time series and random effect, which emphasizes the legitimacy of the two forms of intraclass correlation coefficients. If we overlook the forms of the intraclass correlations, $\rho_{AR(1)}$ is simplified to an overall intraclass correlation coefficient, and functions equivalently to $\rho_{1,ID}$.

The average intraclass correlation of the repeated time measures per individual is $\bar{\rho}$. The current study defines $\bar{\rho}$ as intraclass autocorrelation coefficient (IAC) and $\rho_{1,AR(1)}$ as intraclass correlation coefficient of random effects (ICR). Unlike the ID model that has only one intraclass correlation coefficient (i.e., ICR), the AR(1) model has two forms of intraclass correlations of IAC and ICR, which highlights the serially correlated features of longitudinal data discussed at the beginning.

The ICR is the conventional intraclass correlation, which is

$$\rho_{1,AR(1)} = \frac{\sigma^{(i)}_{AR(1)}}{\sigma^2_{total}} = 1 - \frac{\sigma^{(t)}_{AR(1)}}{\sigma^2_{total}}.$$

Further, since $\rho_{1,ID} = \frac{\sigma^{(i)}_{ID}}{\sigma^2_{total}}$, the relationship between the individual-level random effects of two models is

$$\sigma^{(i)}_{ID} = \sigma^{(t)}_{AR(1)}\bar{\rho} + \sigma^{(i)}_{AR(1)}.$$

Also, since the total error variance is fixed regardless of the model specification, $\sigma^{(t)}_{AR(1)} + \sigma^{(i)}_{AR(1)} = \sigma^2_{total} = \sigma^{(t)}_{ID} + \sigma^{(i)}_{ID}$ gives that

$$\sigma^{(t)}_{ID} = \sigma^{(t)}_{AR(1)}(1 - \bar{\rho}).$$

That is, the estimated intercept random effect in the ID model is smaller than the one in the AR(1) model, while the estimated time-level random effect in the ID model is larger than the one in the AR(1) model. This formula testifies the patterns detected in Murphy and Pituch (2009). The size of the gaps between the random effects of the two models depends on the size of IAC $\bar{\rho}$. Immediately, the random effects of AR(1) can be derived by the following formula:

$$\sigma^{(t)}_{AR(1)} = \frac{\sigma^{(t)}_{ID}}{1 - \bar{\rho}}$$

and

$$\sigma^{(i)}_{AR(1)} = (\sigma^{(i)}_{ID} + \sigma^{(t)}_{ID}) - \sigma^{(t)}_{AR(1)}.$$

Also,

$$\rho_{1,AR(1)} = \frac{\sigma^{(i)}_{AR(1)}}{\sigma^2_{total}} = \frac{\sigma^{(i)}_{ID} - \sigma^{(t)}_{AR(1)}\bar{\rho}}{\sigma^2_{total}} = \frac{\rho_{1,ID} - \bar{\rho}}{1 - \bar{\rho}} = 1 + \frac{\rho_{1,ID} - 1}{1 - \bar{\rho}}. \quad (5.2)$$

Consequently, $\rho_{1,AR(1)}$ is smaller than $\rho_{1,ID}$. The degree of differences between these two ICRs is weighted by the IAC $\bar{\rho}$. When $\bar{\rho}$ is zero, the forms of intraclass correlation reduce to ICR-only as $\rho_{AR(1)} = \rho_{1,AR(1)} = \rho_{1,ID}$. In this case, an estimated ID model is the true model. On the other hand, if $\bar{\rho} = 1$ that each of the time measures of the dependent variable of an individual are exactly the same, $\sigma^{(i)}_{ID} = \sigma^{(t)}_{AR(1)} + \sigma^{(i)}_{AR(1)} = \sigma^2_{total} = \sigma^{(t)}_{ID}$, which is equivalent to a time-individual aggregated or single time-point one level analysis.

Finally, a simple form of the unified single column sum of $\boldsymbol{\psi}_{i,AR(1)}$ is $\tau^*_{AR(1)}$, which is the variance estimate index of the coefficient estimate of $Time_{ti}$:

$$Var_{AR(1)}(\gamma_{10}) = \left\{\sum_{k=1}^{M_K}(Time_{ti}{}'\,\boldsymbol{\psi}_{i,AR(1)}Time_{ti})\right\}^{-1} = \sigma^2\,\tau^*_{AR(1)}\left\{\sum_{k=1}^{M_K}(Time_{ti}{}'\,Time_{ti})\right\}^{-1},$$

and

$$\tau^*_{AR(1)} = 1 + (n_t - 1)\rho_{AR(1)} = 1 + (n_t - 1)\left[\left(1 - \rho_{1,AR(1)}\right)\bar{\rho} + \rho_{1,AR(1)}\right]. \quad (5.3)$$

Then, I construct the VOC to measure the variance inflation size of the estimated variance of the coefficient of the time-level predictor when the AR(1) model is underspecified as ID. The construction rationale is the same as in previous chapters and the conventional design effect that the VOC is the ratio of the variance estimate of the AR(1) model and the variance estimate of the ID model, which yields to

$$VIF_{T,T_{ti}}^{(AR(1)-ID)} = \frac{\tau_{AR(1)}}{\tilde{\tau}_{ID}} = \frac{1+(n_T-1)\rho_{\,AR(1)}\rho_{T_{ti},AR(1)}}{1+(n_T-1)\rho_{1,AR(1)}\rho_{\,T_{ti}}} = \frac{1+(n_T-1)\rho_{1,ID}\rho_{T_{ti},AR(1)}}{1+(n_T-1)\rho_{1,AR(1)}\rho_{\,T_{ti}}}, \quad (5.4)$$

where $\tilde{\tau}_{ID}$ is the scaler weight of the variance estimate of the ID model, which only takes into account of the intraclass correlation due to the random effect. $\tau_{AR(1)}$ is the scaler weight of the variance of the AR(1) model, which takes into account of the two forms of the intraclass correlations. This equation holds the same idea as the ones in previous chapters of quantifying the variance estimate bias when the middle and highest cluster are omitted.

Further, $\rho_{\,T_{ti},AR(1)}$ is the intraclass correlation of repeated time measure predictor in the form of an average lag-1 autocorrelation, while $\rho_{\,TM}$ is the average conventional correlation coefficient. Specifically,

$$\rho_{\,T_{ti},AR(1)} = \frac{1}{n_I}\sum_{i=1}^{n_I}\frac{1}{n_T-1}\left[\frac{\sum_{t=2}^{n_T}(TM_{it}-\overline{TM}_i)(TM_{i(t-1)}-\overline{TM}_i)}{\sum_{t=1}^{n_t}(TM_{it}-\overline{TM}_i)^2}\right]$$

and

$$\rho_{\,T_{ti}} = \frac{1}{n_I}\sum_{i=1}^{n_I}\frac{2}{n_T(n_T-1)}\left[\frac{\sum_{t\neq s}^{n_T}(TM_{it}-\overline{TM}_i)(TM_{is}-\overline{TM}_i)}{\sum_{t=1}^{n_t}(TM_{it}-\overline{TM}_i)^2}\right],$$

where n_I is the number of individuals, TM_{it} is a time occasion measure at time *t* of an individual *i*, and $\overline{TM}_i$ is the individual-level mean of the occasion measures. Group-mean centering of the occasion measure in balanced studies does not produce different values of $\rho_{\,TM,AR(1)}$ and $\rho_{\,TM}$.

The above two intraclass correlation measures of predictors are adapted from the ones in Angrist and Pischke (2008) and Montes-Rojas (2016), which distinguish the difference between these two types of intraclass correlation coefficients of predictors. Specifically, the inclusion of

$\rho_{T_{ti},AR(1)}$ is unique for the time-varying predictors in longitudinal data analysis, which specifies the autocorrelation among one time-measure with the one-time-point-later-measure within individuals. In contrast, the intraclass correlation of predictors $\rho_{T_{ti}}$ is a matter of clustering with equal correlation between any pairs of time-measure within an individual, due to the nature of the model specification of ID. Therefore, in general, $\rho_{T_{ti}}$ is $\frac{2}{n_T}$ times smaller than $\rho_{T_{ti},AR(1)}$ if there are more than two time-measure. If we ignore these two measures of time-varying predictors' intraclass correlation coefficients, $VOC_{T,T_{ti}}^{(AR(1)-ID)}$ tends to be smaller than it should be. In Chapter 2 and 4 for omitted middle and higher cluster levels in non-longitudinal data cases, the intraclass correlation coefficients of a predictor in the denominator and numerator are canceled out since they both equal to the conventional correlation coefficient of the same predictor.

As shown above, the standard error estimates of the time-varying predictors' coefficient are downwardly estimated by the omitted autocorrelation. In contrast, the individual level time-invariant predictor coefficient's standard error is only affected by the overall dependencies, with no need of distinguishing serial correlation or random effects. In other words, the standard error of the individual level time-invariant predictors does not need adjustments in the estimated ID model. The following equation and further simulation results evidence this point.

$$VOC_{T,X_i}^{(AR(1)-ID)} = \frac{\tau_{AR(1)}}{\tilde{\tau}_{ID}} = \frac{1+(n_T-1)\rho_{AR(1)}\rho_{X_i}}{1+(n_T-1)\rho_{ID}\rho_{X_i}} = 1. \quad (5.5)$$

Different from the previous $VIF_{T,T_{ti}}^{(AR(1)-ID)}$ as shown in Eq. 5.4, the denominator of $VIF_{T,X_i}^{(AR(1)-ID)}$ comes from the estimated model that captures all the dependencies, whereas the sources of dependencies are not recognized. Since the predictor of interest X_i here is at the

cluster level and time-invariant, its' forms of intraclass correlations does not need to be distinguished, as long as the overall error variance-covariance are captured. The intraclass correlation of a cluster-level predictor is one (i.e., $\rho_{X_i} = 1$) and canceled out.

Table 5.1 A summary of VOCs when the serial correlation is omitted

Two-level HLM $\boldsymbol{R = ID}$			**Single-level OLS Estimation**		
Level	*Predictor*	*Variance adjustment*	*Level*	*Predictor*	*Variance adjustment*
Time	Time-varying T_{ti}	$VOC_{T,T_{ti}}^{(AR(1)-ID)} > 1$	Time-Student	Time-varying T	/
Individual	Time-invariant X_i	$VOC_{T,X_i}^{(AR(1)-ID)} = 1$		Time-invariant X	$VOC_{T,X_i}^{(AR(1)-OLS)} > 1$

However, when the clustering structure is also omitted and a single-level analysis using OLS estimation is conducted, the standard error estimate of X_i's coefficient $\tilde{\gamma}_{01}$ then needs to be adjusted by the square root of

$$VOC_{T,X_i}^{(AR(1)-OLS)} = \tau_{AR(1)} = 1 + (n_t - 1)\rho_{AR(1)} = 1 + (n_t - 1)\rho_{1,ID}. \quad (5.6)$$

In essence, $VOC_{T,X_i}^{(AR(1)-OLS)}$ shows the sources of the dependencies through $\rho_{AR(1)}$, which is a function of $\rho_{1,AR(1)}$ and $\bar{\rho}$ that have shown in Eq. 5.1. Moreover, if there is no clustering issue (i.e., random effect variance is null) but only autocorrelation, then $VOC_{T,X_i}^{(AR(1)-OLS)}$ reduces to $1 + (n_t - 1)\bar{\rho}$, which mimics the design-based approaches (e.g., DEFF and MF to solve the classic situation of omitting serial correlation in the OLS estimation. Table 5.1 shown above summarizes when and which predictor needs VOC adjustment.

5.3.3 Simulation results

To show the magnitude and direction of estimation bias when R is misspecified, and to examine the performance of the derived VOC formulas, a simulation study is designed with 12 condition sets for the three models of the true AR (1), estimated ID, and estimated single-level models. The conditions are set by the two parameters of the VOC formulas: the number of the time repeated measures n_t of 6, 10, and 30, and the autocorrelation ρ of 0.9, 0.7, 0.5, and 0.2. The total number of individuals n_I, and the true variances of random effects $\sigma^{(t)}_{AR\,(1)}$ and $\sigma^{(i)}_{AR\,(1)}$ are fixed to be 500, 144, and 64 respectively, where the true ICR $\rho_{1,AR(1)}$ is 0.3.

The numbers of the time repeated measures are chosen based on the representative cases in empirical research. For example, the periodicity of ECLS-K: 2011 survey measures are from the kindergarten to the fifth grade that $n_t = 6$. In another example of a daily diary study, the occasion measures can be many more, such as 2 times a day for a half month that $n_t = 30$ (e.g., Ilies & Judge, 2004). The combination of the extensive time measures and relatively smaller autocorrelation gives extremely small average autocorrelation $\bar{\rho}$ values that can be null. These extreme cases serve to prove that, under such circumstances, variance adjustments are not necessary. For each condition, replications of 500 are generated.

Like the earlier discussed simulation studies, the index of relative bias is computed to measure the magnitude of the estimation bias:

$$\text{R. B.}_{est} = \frac{\tilde{\theta} - \theta}{\theta} = \frac{\tilde{\theta}}{\theta} - 1,$$

where θ represents the true parameters from the AR (1) model, including the random effects variances, and standard errors of the repeated time measure $Time_{ti}$ and the individual-level

predictor X_i. Correspondingly, $\tilde{\theta}$ represents the estimates from the estimated ID models. Falsely estimated models lead R. B.$_{est}$ to deviate from zero.

Similarly, a relative bias index of R. B.$_{adj.est}$ is provided for the estimates adjusted by VOCs. The better performance is of the VOCs and less biased of the adjusted estimates, the closer to zero of R. B.$_{adj.est}$ is. Further, a larger difference between R. B.$_{est}$ and R. B.$_{adj.est}$ proves that a biased estimate is more in need of a VOC adjustment. See Appendix 5.A for the simulation parameter setting and the detailed simulation results.

Bias of the Random Effects and the Adjustment Performance

R. B.$_{est}$ of the residual variance estimate $\tilde{\sigma}^{(t)}$ is consistently negative across all models, and the ones of the individual level random effect variance estimate $\tilde{\sigma}^{(i)}$ are positive. In other words, $\tilde{\sigma}^{(t)}$ is underestimated and $\tilde{\sigma}^{(i)}$ is overestimated. The robustness of $\tilde{\sigma}^{(i)}$ is commonly of interest in explaining the proportion of the between-individual variances of the total variance of the outcome. With larger IAC $\bar{\rho}$ that is omitted, R. B.$_{est}$ of $\tilde{\sigma}^{(i)}$ deviates more from zero. For example, when $\rho = 0.9$ and $n_t = 6$, $\tilde{\sigma}_{ID}^{(i)}$ can be 2.77 times as large as the true $\sigma^{(i)}$. When $\rho = 0.2$ and $n_t = 30$, $\tilde{\sigma}_{ID}^{(i)}$ is almost identical to the $\tilde{\sigma}_{AR}^{(i)}$ since $\bar{\rho}$ is close to zero ($\bar{\rho} = 0.017$).

Noticeable, when n_t is small, a small ρ could still result in considerable bias of the random effects estimation. For example, when $\rho = 0.2$ and $n_t = 6$, $\tilde{\sigma}_{ID}^{(i)}$ is 1.17 times larger than $\tilde{\sigma}_{AR}^{(i)}$ (i.e.R. B.$_{est} = 0.17$). Consequently, the estimated $\rho_{1,ID}$ is always larger than the true $\rho_{1,AR(1)}$ as long as $\bar{\rho}$ is not zero. The adjustments of both $\tilde{\sigma}_{ID}^{(i)}$ and $\tilde{\sigma}_{ID}^{(t)}$ are performed ideally across all conditions, where the relative bias are all close to zero (R. B.$_{adj.est} \leq 0.01$) with minimum variances (var(R. B.$_{adj.est}$) ≤ 0.01). The detected patterns prove that the omitted serial

correlation is falsely taken away by the individual-level random effect from the time-level residual, as well as confirming the previously formulated relationships between the random effects from the AR(1) and ID models.

Bias of the Standard Error Estimate of the Coefficient of $Time_{ti}$ and the Adjustment Performance

If the AR (1) structure is omitted, the estimated standard errors of $Time_{ti}$'s coefficient $\tilde{\gamma}_{10}$ is underestimated, as R. B.$_{est}$ are negative across all models. The magnitude of the underestimation bias rises with the increase of ρ and n_t. Fixing $\rho = 0.9$, the estimated standard error of $\tilde{\gamma}_{10}$ is only one-fifth of the true parameter when $n_t = 30$, and three-fifths of the true one when $n_t = 6$. Moreover, the R. B.$_{est}$ values decreases to zero when $\bar{\rho}$ are closing into zero, such as when $\rho = 0.2$ across all n_t. However, the underestimation bias does not diminish. The estimated standard error of $\tilde{\gamma}_{10}$ can still be one-fifth less than the true one.

In terms of the bias adjustment, when $\bar{\rho}$ is larger than 0.1, $VOC_{T,T_{ti}}^{(AR(1)-ID)}$ performs well since $|\text{R. B.}_{adj.est}| \leq 0.05$ and $\text{var}(\text{R. B.}_{adj.est}) \leq 0.001$, except for the case of when $\rho = 0.5$ and $n_t = 10$ ($\bar{\rho} = 0.178$). The performance of $VOC_{T,T_{ti}}^{(AR(1)-ID)}$ are also relatively better when the occasion measures are not extensive. When ρ is moderate and small (i.e., 0.5 and 0.2), R. B.$_{adj.est}$ tends to be positive, though smaller than 0.1 when n_t is 6 and smaller than 0.3 when n_t is 10. If n_t gets extensively large to be 30, $VOC_{T,T_{ti}}^{(AR(1)-ID)}$ tends to make undesired overcorrections that R. B.$_{adj.est}$ is larger than 0.5, or even as large as 1. Type II error can thus be caused. In these cases, $\bar{\rho}$ are around 0.05 and smaller. The undesired overcorrection pattern could also be related to the values of the intraclass correlations of predictors $\rho_{Time_{ti},AR(1)}$ and $\rho_{Time_{ti}}$.

As shown in $n_t = 30$, $\rho_{T_{ti},AR(1)} = 0.9$ while $\rho_{Time_{ti}} = 0.06$, in which the extremely small $\rho_{Time_{ti}}$ produces an extremely small denominator of $VOC_{T,T_{ti}}^{(AR(1)-ID)}$. As a Result, the corresponding $VOC_{T.T_{ti}}^{(AR(1)-ID)}$ tends to be much larger than it should be.

Bias of the Standard Error Estimate of the Coefficient of X_i and the Adjustment Performance

As expected, when the clustering structure is omitted, there are underestimation issues of the standard error estimates of the individual-level predictor X_i's coefficient $\tilde{\gamma}_{01}$. Consistently across all conditions, R. B.$_{est}$ are negative and around from -0.5 to -0.7. Equivalently, the estimated standard error estimates from the single-level analyses are only half of or even smaller than the true parameter. $VOC_{T,X_i}^{(AR(1)-OLS)}$ performs desirable in all models that R. B.$_{adj.est}$ are close to zero, except for one noticeable overcorrection case of when $n_t = 30$ and $\rho = 0.7$.

5.4 Empirical Example and Sensitivity Analysis

The selected empirical example is in Taylor et al. (2010), which applies two-level linear growth models to examine the impacts of between-student and within-student motivational regulations and psychological needs on three motivational outcome of effort, intentions, and physical activity growth. The 178 participant students come from an England school who are in grade-level 6 through 10. The three outcome variables are measured from three semesters' surveys.

The original study does not specify the time-level random effect variance structure, thus, assuming an AR (1) structure is underspecified as ID, the current study presents examples of utilizing the sensitivity analysis to test the robustness of the time-varying predictors. The employed models are the ones with outcome predictor of students' intentions to exercise (see

Table 1 and 2 of Taylor et al., 2010 for the detailed model reports). In both models, the random slopes are not significant, and the variance estimates are close to 0 (i.e., 0.1 and 0.01 respectively). Thus, I use the above VOC formulas that are initially constructed from the random intercept models as an elementary example. Following the suggested steps of conducting the sensitivity analysis in the heuristics diagram of Figure 3.3 of Chapter 3, the threshold VOC is calculated at first. Then examples that are minimum and maximum IAC values are presented to show the boundaries of robustness.

Table 5.2 Sensitivity analysis of the time-varying predictor: competence

	ID	**AR (1)**				$\hat{\beta}$	0.27		
$\rho_{C_{ti}}$	0.53	0.79				$StE^{\#}$	0.14	$t^{\#}$	1.96
		IAC=0.10	*IAC=0.51*	*IAC=0.30*		StE_{ID}	0.13	t_{ID}	2.08
$\sigma^{(i)}$	1.20	1.08	0.02	0.72					
$\sigma^{(t)}$	1.12	1.24	2.30	1.60					
ICR: ρ_1	0.52	0.46	0.01	0.31					
IAC: $\bar{\rho}$		**VOC Index**	$\sqrt{VOC}$	StE_{voc}	W_{OC} & ES_{OC}	R_{OC}			
	0.00	Threshold VOC_0	1.060	0.138	*NA*	*NA*			
$\bar{\rho}_{min}$	0.10	VOC_{min}	1.105	0.144	0.095	0.265			
$\bar{\rho}_{max}$	0.51	VOC_{max}	1.341	0.174	0.254	0.719			
	0.30		1.170	0.152	0.145	0.508			

Table 5.2 above presents the sensitivity analysis results of the time-varying predictor competence in the first model. The right upper corner shows that the robustness of the competence predictor is not desirable since the t statistic is 2.08, which is almost at the threshold $t^{\#}$ of 1.96. Therefore, any small serial correlation can lead to a Type I error. In this case, the threshold VOC is less useful. In the table, the grey cells are fixed values, including parameters

that are provided in the original study, and the ones that are set to achieve minimum and maximum IAC values.

Setting a minimum IAC being 0.1, the corresponding square root of VOC is 1.105, and the ICR of the AR (1) model (i.e., $\mathrm{ICR}_{\mathrm{AR(1)}}$) is close to the ICR of the ID model (i.e., $\mathrm{ICR}_{\mathrm{ID}}$). The original table provides the intraclass correlation of the predictor competition ($\rho_{C_{ti},AR(1)}$) being 0.79, and thus $\rho_{C_{ti}}$ turning to be 0.53. In this setting, the robustness of inference (i.e., W_{OC}) or the effect size (i.e., EF_{OC}) reduces by 1%, and the risk of making Type I error increases by 26.5%. When setting a minimum $\mathrm{ICR}_{\mathrm{AR(1)}}$ being 0.01, the corresponding square root of VOC is 1.341, and the IAC is 0.51. This setting offers the upper bound of the possible IAC and the magnitude of bias. The robustness of inference or the effect size reduces by 25.4 %, and the risk of making Type I error increases by 71.9 %. These two settings have correspondingly lag-1 autocorrelation ρ values of 0.15 and 0.7, which form a reasonable boundary of a potentially omitted serial correlation.

Tables 5.3 and 5.4 show the sensitivity analysis examining the time-varying predictors of intrinsic regulation and external regulation in the second model. For both predictors, they have the same IAC values since they share the same random effects in the same model, while they have different VOCs due to the different intraclass correlation of predictors. Provided by the original study, $\rho_{IR_{ti},AR(1)} = 0.73$ and $\rho_{ER_{ti},AR(1)} = 0.53$. The estimated effects of these two predictors are relatively robust. Specifically, their threshold VOCs are larger than the upper bound of possible VOCs. Thus, no risk of Type I error issue emerges.

Table 5.3 Sensitivity analysis of the time-varying predictor: intrinsic regulation

	ID	**AR (1)**				$\hat{\beta}$	0.36		
$\rho_{IR_{ti}}$	0.49	0.73				$StE^{\#}$	0.18	$t^{\#}$	1.96
		IAC=0.10	*IAC=0.66*	*IAC=0.30*		StE_{ID}	0.10	t_{ID}	3.60
$\sigma^{(i)}$	1.61	1.52	0.02	1.26					
$\sigma^{(t)}$	0.81	0.90	2.40	1.16					
ICR: ρ_1	0.67	0.63	0.01	0.52					
IAC: $\bar{\rho}$		VOC Index	$\sqrt{VOC}$	StE_{voc}	W_{OC} & ES_{OC}	R_{OC}			
	0.00	Threshold VOC_0	1.837	0.184	NA	NA			
$\bar{\rho}_{min}$	0.10	VOC_{min}	1.106	0.111	0.096	NA			
$\bar{\rho}_{max}$	0.66	VOC_{max}	1.397	0.140	0.284	NA			
	0.30		1.143	0.114	0.125	NA			

Table 5. 4 Sensitivity analysis of the time-varying predictor: external regulation

	ID	**AR (1)**				$\hat{\beta}$	0.35		
$\rho_{ER_{ti}}$	0.35	0.53				$StE^{\#}$	0.18	$t^{\#}$	1.96
		IAC=0.10	*IAC=0.66*	*IAC=0.30*		StE_{ID}	0.08	t_{ID}	4.38
$\sigma^{(i)}$	1.61	1.52	0.02	1.26					
$\sigma^{(t)}$	0.81	0.90	2.40	1.16					
ICR: ρ_1	0.67	0.63	0.01	0.52					
IAC: $\bar{\rho}$		VOC Index	$\sqrt{VOC}$	StE_{voc}	W_{OC} & ES_{OC}	R_{OC}			
	0.00	Threshold VOC_0	2.232	0.179	NA	NA			
$\bar{\rho}_{min}$	0.10	VOC_{min}	1.087	0.087	0.080	NA			
$\bar{\rho}_{max}$	0.66	VOC_{max}	1.301	0.104	0.232	NA			
	0.30		1.116	0.089	0.104	NA			

However, the robustness of inference and effect size still needs attention. With a maximum IAC $\bar{\rho}_{max} = 0.66$, the robustness of inference and effect size reduces by 28 % and 23 %, respectively, of the predictors of intrinsic regulation and external regulation. With a minimum IAC $\bar{\rho}_{min} = 0.10$, the robustness of inference and effect size reduces by around 1 % for both predictors.

In sum, the sensitivity analysis provides that the inferences made for the regulation predictors are relatively strong if the AR (1) structure is omitted. However, the inference made for competence needs attention because even a minimum omitted autocorrelation can lead to serious Type I error issue. This evidence is critical since the conclusion drawing on the within-student level competence is the focus of the original study.

5.5 Conclusion and Future Research

Consistent with previous research (Alejo, et al., 2018; Betrand et al., 2004), the current study proves that when the chronological order structure within cluster units are omitted in multilevel analysis for longitudinal data, the intraclass correlation due to individual-level random effect variance takes over the serial correlation. To the author's knowledge, the current study is the first that formulated this relationship of random effects and serial correlations when R is underspecified from AR (1) to ID. This study further determines that the magnitude of the overestimation of the individual-level random effect variance is weighed by the IAC $\bar{\rho}$. The conceptualization of IAC and ICR provides new understandings of the conventional intraclass correlation coefficient, and evidence to decide which level's predictors are essentially needed.

Further, the derivations of VOCs are conducted separately for time-level and individual-level predictors. These formulas produce consistent suggestions with the simulation-based findings from the earlier discussed prior research, such as Ferron et al. (2002) and LeBeau (2018). Specifically, when the true AR(1) is completely omitted, time-varying predictors need adjustments, while the time-invariant predictors do not. Noticeably, the current study does not recommend adjusting the standard error estimates of fixed effects when the occasion measures are extensive, and the hypothesized lag-1 autocorrelation is small such that the IAC is smaller

than 0.2. Employing the sensitivity analysis framework developed in Chapter 3, empirical researchers and readers are able to easily find evidence of the extent to which the inference is robust. The strategies are provided with an empirical research example (i.e., Taylor, et al., 2010).

The current study sets models with only random intercept. However, random slopes are common in longitudinal data analysis. Particularly, if the random effect of slopes is ignored in modeling, Type I error rate inflates (LeBeau, 2018). Including the random slopes in the current study increases the complexity of the variance-covariance structure, since the covariance units depend on the occasion measures. This complexity can be addressed in future studies. For instance, with the experience of constructing an averaged autocorrelation parameter (i.e., IAC) for the descending serial correlation pattern, an average covariance parameter can be similarly constructed, as long as the overall error variance-covariance are captured correctly. However, the precision and consistence of the averaged autocorrelation and covariance parameters could be affected by the missing data and unbalanced design, which need further tests.

Also, the current study only explored the relationship between the ID and AR (1) structures. In future studies, the interrelationship between other commonly used alternative R structures will be developed. For example, AR (1) is easily to relate to ARMA (1,1). Finally, future research may study the omitted serial correlation in three-level models. For instance, a higher cluster level of school could exist. Comparing with the current study, two additional intraclass correlations emerge: the ICR and IAC that are school specific (Alejo, et al., 2018). Then the quantification of estimation bias due to omitted serial correlation are complex as to distinguish the sources of the intraclass correlations.

APPENDICES

APPENDIX 2A

Intraclass Correlation Coefficients in A Three-Level Model

In the current study, the intraclass correlation coefficients (ICCs) of classroom- and school-level are defined as $\rho_1 = \frac{\sigma^{(k)}+\sigma^{(j)}}{\sigma^2}$ and $\rho_2 = \frac{\sigma^{(k)}}{\sigma^2}$. Another commonly used definition of ICCs is $\rho_1^* = \frac{\sigma^{(j)}}{\sigma^2}$ and $\rho_2 = \frac{\sigma^{(k)}}{\sigma^2}$. The distinction of these two methods of ICCs occurs only at ρ_1 and ρ_1^*. Hox, Moerbeek, and Van de Schoot (2010) summarized that these two methods are both correct, though having slightly different focuses. The latter method has a focus on decomposing the variance from each level that ρ_1^* identifies the unique classroom-level variance.

In the first method, ρ_1 is derived as the following

$$\rho_1 = corr(y_{ijk}, y_{i'jk}) = \frac{cov(y_{ijk} - \hat{y}_{ijk}, y_{i'jk} - \hat{y}_{i'jk})}{sd(y_{ijk} - \hat{y}_{ijk}) * sd(y_{i'jk} - \hat{y}_{i'jk})},$$

where the denominator is the total error variance σ^2, and the numerator is:

$$cov(y_{ijk} - \hat{y}_{ijk}, y_{i'jk} - \hat{y}_{i'jk}) = cov(u_{00k} + r_{0jk} + \varepsilon_{ijk}, u_{00k} + r_{0jk} + \varepsilon_{i'jk})$$

$$= cov(u_{00k} + r_{0jk}, u_{00k} + r_{0jk}) + cov(\varepsilon_{ijk}, \varepsilon_{i'jk})$$

$$+cov(u_{00k} + r_{0jk}, r_{0jk}) + cov(u_{00k} + r_{0jk}, \varepsilon_{i'jk}).$$

With assumptions of random effects having zero covariance with each other,

$$cov(y_{ijk} - \hat{y}_{ijk}, y_{i'jk} - \hat{y}_{i'jk}) = cov(u_{00k}, u_{00k}) + cov(r_{0jk}, r_{0jk}) + 2cov(u_{00k}, r_{0jk})$$

$$= var(u_{00k}) + var(r_{0jk}) = \sigma^{(k)} + \sigma^{(j)}.$$

As shown, $\rho_1 = \frac{\sigma^{(k)}+\sigma^{(j)}}{\sigma^2}$ measures the expected correlation between two students who are in the same class and, also, in the same school. Conversely, $\rho_2 = \frac{\sigma^{(k)}}{\sigma^2}$ measures the expected correlation between two students who are in the same school but from different classes.

APPENDIX 2B

A Summary of Model Specification, Assumption, and Estimation

Table 2B.1 Summary of model specification, assumption, and estimation contrasting the two-level estimated model omitting the middle cluster level and the three-level satisfactory model.

Model Specification, Assumption, and Estimation	**Two-level Estimated Model**	**Three-level Satisfactory Model**
1. Multi-stage Sampling Design and Experimental Design with Clusters.	Multi-stage Sampling 1) In a three-stage sampling where PSUs are schools, SSUs are classrooms, and USUs are students, the middle classroom **deliberate** cluster level is omitted in modeling. 2) Or, in a two-stage sampling design where PSUs are schools and SSUs are students, the middle classroom **incidental** cluster level is omitted in modeling.	Multi-stage Sampling 1) The model corresponds with the three-stage sampling design that all the sampling stages as deliberate cluster levels are specified in modeling. 2) Or, in a two-stage sampling design where PSUs are schools and SSUs are students, the middle classroom **incidental** cluster level is included in modeling.
	RCT: treatment is randomly assigned to schools.	RCT: treatment is randomly assigned to schools.
2. All relevant predictors are included in the model.	Predictors of interest to answer research questions: 1) $X_{i(j)k}$ is a student-level predictor. 2) $W_{i(j)k}$ is a (falsely disaggregated) student-level predictor. 3) $Z_{(j)k}$ is a school-level predictor.	Predictors of interest to answer research questions: 1) X_{ijk} is a student-level predictor. 2) W_{jk} is a classroom-level predictor. 3) Z_k is a school-level predictor.
	Relevant covariates, such as contextual factors, at each level based on subject-matter knowledge.	Relevant covariates, such as contextual factors, at each level based on subject-matter knowledge.

Table 2B.1 (cont'd)

3. Random intercept only.	1) Schools differ across the average value of outcomes. 2) The slopes at all levels do not differ across schools.	1) Schools and classrooms within schools differ across the average value of outcomes. 2) The slopes at all levels do not differ across schools.
4. The error variance covariance structure is properly specified.	$\widetilde{\boldsymbol{\psi}}_K = \sigma^2\left[(1-\rho)\boldsymbol{I} + \rho\, \boldsymbol{l}_{n_0}\boldsymbol{l'}_{n_0}\right]$.	$\boldsymbol{\psi}_K = \sigma^2\{\boldsymbol{I_K} \otimes \left[(1-\rho_1)\boldsymbol{I}_J + (\rho_1-\rho_2)\boldsymbol{l}_J\boldsymbol{l'}_J\right] + \rho_2\, \boldsymbol{l}_{n_0}\boldsymbol{l'}_{n_0}\}$.
	Parameters: 1) One ICC: ρ measures the similarity of students within the same school *k*, regardless of classrooms. 2) One cluster size: n_0 is the average number of students within a school *k*.	Parameters: 1) Two ICCs: ρ_1 is the expected correlation of two randomly drawn students from the same classroom *j* in a school *k*, and ρ_2 is the expected correlation of two randomly drawn students from the same school *k*. 2) Two cluster size: n_L is the average class size and n_H is the average number of teachers within each school *k*.
5. The within-cluster residuals follow a multivariate normal distribution.	$\tilde{\varepsilon}_{ik} \sim N\left(0, \tilde{\sigma}^{(i)}\right)$, where $\tilde{\sigma}^{(i)}$ is conditioned on predictors and covariates.	$\varepsilon_{ijk} \sim N\left(0, \sigma^{(i)}\right)$, where $\sigma^{(i)}$ is conditioned on predictors and covariates.
6. The random effects follow a multivariate normal distribution.	$\tilde{u}_{0k} \sim N\left(0, \tilde{\sigma}^{(k)}\right)$, where $\tilde{\sigma}^{(k)}$ is conditioned on predictors and covariates. And, the group effects $\tilde{u}_{0k}$ is independent and identically distributed that no higher cluster level exists.	$r_{0jk} \sim N\left(0, \sigma^{(j)}\right)$ and $u_{00k} \sim N\left(0, \sigma^{(k)}\right)$, where $\sigma^{(j)}$ and $\sigma^{(k)}$ are conditioned on predictors and covariates. And, the group effects u_{0k} is independent and identically distributed that no higher cluster level exists.
7. Homoscedasticity.	1) Constant error variance of all levels conditioned on predictors. 2) Or corrected heteroskedastic patterns for the specified nesting structure.	1) Constant error variance of all levels conditioned on predictors. 2) Or corrected heterogeneity for the specified nesting structure. 3) The assumptions still hold after including the omitted cluster level.

Table 2B.1 (cont'd)

8. The within-cluster residuals and the random effects do not covary.	$cov(\tilde{\varepsilon}_{ik}, \tilde{u}_{0k}) = 0.$	$cov(\varepsilon_{ijk}, r_{0jk}) = 0,$ $cov(\varepsilon_{ijk}, u_{00k}) = 0,$ and $cov(r_{0jk}, u_{00k}) = 0.$
9. The predictors do not covary with the residuals and random effects at any other level.	1) X_{ik} and W_{ik} are group-mean centered. 2) Assume no omitted confounding variables at all levels.	1) X_{ijk} and W_{jk} are group-mean centered. 2) Assume no omitted confounding variables at all levels.
10. Sample size.	A sufficient large sample size (both cluster numbers and cluster size) at all levels to satisfy the desired power and for the asymptotic inference.	A sufficient large sample size (both cluster numbers and cluster size) at all levels to satisfy the desired power and for the asymptotic inference.
	Balanced design or at least almost equal sample size of clusters.	Balanced design or at least almost equal sample size of clusters.
11. Estimation.	1) (Restricted) Maximum Likelihood. 2) Design-based approach for the standard error bias correction.	1) (Restricted) Maximum Likelihood.

Note. The listed model specification, assumption, and estimation in the first column are summarized from McNeish and Kelley, (2019, p. 26), McNeish et al. (2016, p. 116), Snijders and Berkhof (2008) and Snijders & Bosker, (2012, p. 102).

APPENDIX 2C

Simulation Parameter Settings and Result of VOCs of Omitting the Middle Cluster Level

Table 2C.1 Simulation parameter settings.

$\sigma^{(j)}$ or ρ_0	$\sigma^{(k)}$ or ρ_2	$\sigma^{(i)}$	n_L Avg. Class size	n_H Avg. No. of teachers/classrooms per school	η	$\tilde{\sigma}^{(k)}$ or ρ	$\tilde{\sigma}^{(i)}$
0.2	0.2	0.6	5	10	0.08	0.22	0.78
0.5	0.2	0.3				0.24	0.76
0.7	0.2	0.1				0.26	0.74
0.2	0.7	0.1				0.72	0.28
0.2	0.2	0.6	10	5	0.18	0.24	0.76
0.5	0.2	0.3				0.29	0.71
0.7	0.2	0.1				0.33	0.67
0.2	0.7	0.1				0.74	0.26
0.2	0.2	0.6	25	2	0.49	0.30	0.70
0.5	0.2	0.3				0.45	0.55
0.7	0.2	0.1				0.54	0.46
0.2	0.7	0.1				0.80	0.20

Table 2C.2 Relative bias of estimates of variances when $\sigma^j = \rho_0 = 0.2$, $\sigma^k = \rho_2 = 0.2$.

Parameter	$\mathbf{n}_L$	$\boldsymbol{\eta}$	**R. B.$_{est}$ of HLM**		**R. B.$_{adj.est}$ of HLM**	
			Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Residual variance $\tilde{\sigma}^{(i)}$	5	0.08	0.30 (0)	[0.13, 0.47]	0 (0)	[0, 0]
	10	0.18	0.27 (0)	[0.13, 0.47]	0 (0)	[0, 0]
	25	0.49	0.17 (0)	[0.10, 0.26]	0 (0)	[0, 0.07]
School-level random effect variance $\tilde{\sigma}^{(k)}$	5	0.08	0.11 (0)	[0.03, 0.45]	0 (0)	[0, 0]
	10	0.18	0.27 (0.05)	[0.06, 2.06]	0 (0)	[0, 0]
	25	0.49	0.39 (0.19)	[1.03, 0.12]	0 (0)	[0, 0]
Standard error of $X_{i(j)k}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	0.08 (0)	[0.05, 0.10]	-0.06 (0)	[-0.10, -0.04]
	10	0.18	0.09 (0)	[0.06, 0.13]	-0.03 (0)	[-0.05, -0.02]
	25	0.49	0.07 (0)	[0.04, 0.11]	-0.01 (0)	[-0.04, 0.00]
Standard error of $W_{i(j)k}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.30 (0)	[-0.36, -0.18]	0.13 (0)	[0.04, 0.24]
	10	0.18	-0.45 (0)	[-0.53, -0.34]	0.16 (0)	[0.06, 0.34]
	25	0.49	-0.63 (0)	[-0.73, -0.44]	0.15 (0.01)	[-0.05, 0.42]
Standard error of $Z_{(j)k}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	0 (0)	[0, 0]	0 (0)	[0, 0]
	10	0.18	0 (0)	[0, 0]	0 (0)	[0, 0]
	25	0.49	0 (0)	[-0.16, 0]	0 (0)	[-0.01, 0]

Table 2C.2 (cont'd)

Parameter	n_L	η	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	0.20 (0)	[0.14, 0.30]	-0.07 (0)	[-0.13, -0.04]
	10	0.18	0.24 (0)	[0.17, 0.31]	-0.04 (0)	[-0.08, -0.01]
	25	0.49	0.26 (0)	[0.17, 0.36]	-0.02 (0)	[-0.10, 0.02]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.21 (0)	[-0.32, -0.10]	0.06 (0)	[-0.01, 0.13]
	10	0.18	-0.38 (0)	[-0.50, -0.23]	0.04 (0)	[-0.01, 0.10]
	25	0.49	-0.56 (0)	[-0.70, -0.36]	0.01 (0)	[-0.06, 0.11]
Standard error of $Z_{(jk)}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	-0.68 (0)	[-0.77, -0.49]	-0.01 (0)	[-0.09, 0.12]
	10	0.18	-0.70 (0)	[-0.78, -0.50]	-0.01 (0)	[-0.09, 0.11]
	25	0.49	-0.73 (0)	[-0.80, -0.58]	-0.02 (0)	[-0.24, 0.12]

Table 2C.3 Relative bias of estimates of variances when $\sigma^j = \rho_0 = 0.5$, $\sigma^k = \rho_2 = 0.2$.

Parameter	n_L	η	R. B.$_{est}$ of HLM		R. B.$_{adj.est}$ of HLM	
			Mean (Variance)	Range	Mean (Variance)	Range
Residual variance $\tilde{\sigma}^{(i)}$	5	0.08	1.52 (0.04)	[0.95, 2.08]	0 (0)	[0, 0]
	10	0.18	1.36 (0.07)	[0.77, 2.16]	0 (0)	[0, 0.04]
	25	0.49	0.81 (0.07)	[0.28, 1.84]	0.01 (0)	[-0.01, 0.37]
School-level random effect variance $\tilde{\sigma}^{(k)}$	5	0.08	0.31 (0.06)	[0.09, 2.50]	0 (0)	[0, 0]
	10	0.18	0.39 (0.13)	[0.13, 0.82]	0 (0)	[0, 0]
	25	0.49	0.81 (0.43)	[0.13, 2.47]	0 (0)	[0, 0]
Standard error of $X_{i(j)k}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	0.45 (0)	[0.38, 0.54]	-0.09 (0)	[-0.35, 0.00]
	10	0.18	0.47 (0)	[0.38, 0.59]	-0.04 (0)	[-0.20, 0.06]
	25	0.49	0.34 (0)	[0.22, 0.49]	-0.01 (0)	[-0.32, 0.09]
Standard error of $W_{i(j)k}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.48 (0)	[-0.50 -0.44]	0.01 (0)	[-0.05, 0.08]
	10	0.18	-0.63 (0)	[-0.66 -0.59]	-0.01 (0)	[-0.09, 0.06]
	25	0.49	-0.78 (0)	[-0.82 -0.71]	-0.13 (0.01)	[-0.27 0.03]
Standard error of $Z_{(j)k}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	0 (0)	[0, 0]	0 (0)	[0, 0]
	10	0.18	0 (0)	[-0.07, 0]	0 (0)	[0, 0]
	25	0.49	-0.01 (0)	[-0.21, 0.01]	0 (0)	[-0.01, 0]

Table 2C.3 (cont'd)

Parameter	n_L	η	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	0.65 (0)	[0.54, 0.81]	-0.09 (0)	[-0.39, 0.01]
	10	0.18	0.73 (0)	[0.60, 0.85]	-0.05 (0)	[-0.23, 0.08]
	25	0.49	0.78 (0)	[0.59, 1.02]	-0.01 (0)	[-0.44, 0.14]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.40 (0)	[-0.44, -0.37]	0.03 (0)	[-0.03, 0.10]
	10	0.18	-0.57 (0)	[-0.63, -0.46]	<0.01 (0)	[-0.13, 0.17]
	25	0.49	-0.71 (0)	[-0.77, -0.66]	<0.01 (0.01)	[-0.09, 0.11]
Standard error of $Z_{(jk)}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	-0.70 (0)	[-0.78, -0.51]	-0.01 (0)	[-0.12, 0.12]
	10	0.18	-0.73 (0)	[-0.80, -0.59]	-0.01 (0)	[-0.19, 0.16]
	25	0.49	-0.78 (0)	[-0.83, -0.72]	-0.04 (0.01)	[-0.28, 0.19]

Table 2C.4 Relative bias of estimates of variances when $\sigma^j = \rho_0 = 0.7$, $\sigma^k = \rho_2 = 0.2$.

Parameter	n_L	η	R. B.$_{est}$ of HLM		R. B.$_{adj.est}$ of HLM	
			Mean (Variance)	Range	Mean (Variance)	Range
Residual variance $\tilde{\sigma}^{(i)}$	5	0.08	6.38 (0.55)	[4.19, 8.69]	0 (0)	[0, 0]
	10	0.18	5.76 (0.87)	[3.43, 8.74]	0.01 (0)	[-0.01, 0.41]
	25	0.49	3.36 (1.17)	[1.15, 8.13]	0.10 (0.08)	[-0.37, 1.99]
School-level random effect variance $\tilde{\sigma}^{(k)}$	5	0.08	0.46 (0.22)	[-1.00, 6.64]	0 (0)	[-0.02, 0]
	10	0.18	0.58 (0.18)	[1.07, 0.21]	0 (0)	[0, 0]
	25	0.49	1.00 (0.50)	[2.83, 0.20]	0 (0)	[0, 0]
Standard error of $X_{i(j)k}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	1.47 (0)	[1.30, 1.66]	-0.10 (0.04)	[-0.92, 0.28]
	10	0.18	1.47 (0.01)	[1.26, 1.74]	-0.03 (0.06)	[-0.99, 0.45]
	25	0.49	1.09 (0.01)	[0.78, 1.46]	0.03 (0.08)	[-0.94, 0.45]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.55 (0)	[-0.55, -0.53]	0.02 (0)	[-0.03, 0.08]
	10	0.18	-0.69 (0)	[-0.70, -0.68]	-0.08 (0)	[-0.16, 0.01]
	25	0.49	-0.83 (0)	[-0.85, -0.81]	-0.24 (0.01)	[-0.34, -0.14]
Standard error of $Z_{(jk)}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	0 (0)	[0, 0]	0 (0)	[0, 0]
	10	0.18	0 (0)	[-0.98, 0]	0 (0)	[-0.01, 0]
	25	0.49	-0.01 (0)	[-0.30, 0.04]	0 (0)	[-0.01, 0]

Table 2C.4 (cont'd)

Parameter	n_L	η	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	1.83 (0.01)	[1.61, 2.13]	-0.11 (0.04)	[-0.93, 0.29]
	10	0.18	1.99 (0.01)	[1.68, 2.24]	-0.03 (0.06)	[-0.99, 0.50]
	25	0.49	2.07 (0.02)	[1.65, 2.59]	0.05 (0.10)	[-0.94, 0.58]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.48 (0)	[-0.53, -0.40]	0.01 (0)	[-0.12, 0.16]
	10	0.18	-0.63 (0)	[-0.67, -0.55]	0.01 (0)	[-0.09, 0.08]
	25	0.49	-0.75 (0)	[-0.79, -0.61]	<0.01 (0.01)	[-0.14, 0.12]
Standard error of $Z_{(jk)}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	-0.71 (0)	[-1.00, -0.55]	-0.01 (0)	[-0.99, 0.15]
	10	0.18	-0.74 (0)	[-0.99, -0.66]	-0.02 (0)	[-0.97, 0.21]
	25	0.49	-0.81 (0)	[-0.84, -0.78]	-0.05 (0.01)	[-0.38, 0.21]

Table 2C.5 Relative bias of estimates of variances when $\sigma^j = \rho_0 = 0.2$, $\sigma^k = \rho_2 = 0.7$.

Parameter	n_L	η	R. B.$_{est}$ of HLM		R. B.$_{adj.est}$ of HLM	
			Mean (Variance)	Range	Mean (Variance)	Range
Residual variance $\tilde{\sigma}^{(i)}$	5	0.08	1.82 (0.05)	[1.16, 2.48]	0 (0)	[0, 0]
	10	0.18	1.63 (0.09)	[0.93, 2.58]	0 (0)	[0, 0]
	25	0.49	0.97 (0.12)	[0.24, 2.20]	0 (0)	[0, 0.05]
School-level random effect variance $\tilde{\sigma}^{(k)}$	5	0.08	0.03 (0)	[0.01, 0.09]	0 (0)	[0, 0]
	10	0.18	0.07 (0)	[0.02, 0.24]	0 (0)	[0, 0]
	25	0.49	0.18 (0.01)	[0.03, 0.76]	0 (0)	[0, 0]
Standard error of $X_{i(j)k}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	0.54 (0)	[0.45, 0.63]	-0.05 (0.05)	[-0.80, 0.28]
	10	0.18	0.56 (0)	[0.45, 0.69]	-0.02 (0.05)	[-0.82, 0.28]
	25	0.49	0.40 (0)	[0.26, 0.57]	0.01 (0.04)	[-0.77, 0.27]
Standard error of $W_{i(j)k}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.49 (0)	[-0.51, -0.46]	0.08 (0)	[-0.06, 0.29]
	10	0.18	-0.64 (0)	[-0.67, -0.61]	0.07 (0)	[-0.07, 0.24]
	25	0.49	-0.79 (0)	[-0.83, -0.69]	-0.06 (0)	[-0.23, 0.16]
Standard error of $Z_{(j)k}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	0 (0)	[0, 0]	0 (0)	[0, 0]
	10	0.18	0 (0)	[0, 0]	0 (0)	[0, 0]
	25	0.49	0 (0)	[-0.03, 0]	0 (0)	[-0.01, 0]

Table 2C.5 (cont'd)

Parameter	n_L	η	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	5	0.08	1.84 (0.02)	[1.45, 2.51]	-0.01 (0.08)	[-0.83, 0.51]
	10	0.18	2.00 (0.03)	[1.58, 2.46]	0.03 (0.09)	[-0.85, 0.57]
	25	0.49	2.06 (0.03)	[1.50, 2.66]	0.09 (0.11)	[-0.83, 0.67]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{20}$	5	0.08	-0.05 (0.01)	[-0.20, 0.15]	0.27 (0.00)	[0.08, 0.55]
	10	0.18	-0.33 (0.01)	[-0.55, -0.05]	0.15 (0.02)	[0.01, 034]
	25	0.49	-0.55 (0.01)	[-0.65, -0.43]	0.06 (0)	[-0.13, 0.27]
Standard error of $Z_{(jk)}$ coefficient $\tilde{\gamma}_{01}$	5	0.08	-0.83 (0)	[-0.84, -0.78]	-0.01 (0.01)	[-0.14, 0.09]
	10	0.18	-0.83 (0)	[-0.85, -0.78]	-0.04 (0.01)	[-0.34, 0.27]
	25	0.49	-0.84 (0)	[-0.85, -0.83]	-0.01 (0.02)	[-0.14, 0.11]

APPENDIX 3A

Quantifying the Robustness of Inference with Type 2 Error

In cases of when $t_{ols} < t^{\#}$, Type 2 error may occur. The discussion serves scenarios of when the adjacent higher cluster level is omitted, the lower level cluster's predictors' coefficients have overestimated standard error estimates. For example, Chapter 2 showed that VOCs of the individual level predictor $X_{i(j)k}$ (and $X_{i(jk)}$) are smaller than 1 when the upper middle cluster level is omitted. Further in Chapter 4, the standard error estimate of the middle cluster level predictor $W_{ij(k)}$ coefficient could also be overestimated when the highest cluster level is omitted. This scenario of having potential risks of making Type II error is demonstrated using an empirical study in Chapter 4 with implementing the below robustness inference measures.

Identical to the discussed rationale for comparing the deviation of the estimated models from the true models, Figure 3.A.1 shows the two possible scenarios of having or not having Type II error when the t-statistic is smaller than the t critical value. Unlike Figure 3.2, Δ turns to be deflation instead of inflation. The definitions of quantifying the deviations of the t statistics to the t critical value remain the same, while the formulas are reversed as in Type I error discussions that $\Delta_1 = t_{vif} - t^{\#}$ and $\Delta_2 = t^{\#} - t_{ols}$. In panel (a), Type II error does not occur since $t_{ols} < t_{vif} < t^{\#}$; in Panel (b), Type II error occurs since $t_{ols} < t^{\#} < t_{vif}$.

Following the ideas of constructing the measures of robustness of inference and effect size when $t_{ols} > t^{\#}$ that have been discussed previously, these two measures are adapted for the current setting of $t_{ols} < t_{vif} < t^{\#}$ (i.e., no Type II error) :

$$W_{OC} = \frac{\Delta}{t_{vif}} = \frac{\Delta_2 + \Delta_1}{t_{vif}} = \frac{t_{ols} - t_{vif}}{t_{vif}} = 1 - \frac{StE_{vif}}{StE_{ols}} = 1 - \sqrt{VIF},$$

and

$$ES_{OC} = \frac{ES_{OLS} - ES_{VIF}}{ES_{VIF}} = 1 - \frac{StE_{vif}}{StE_{ols}} = 1 - \sqrt{VIF}.$$

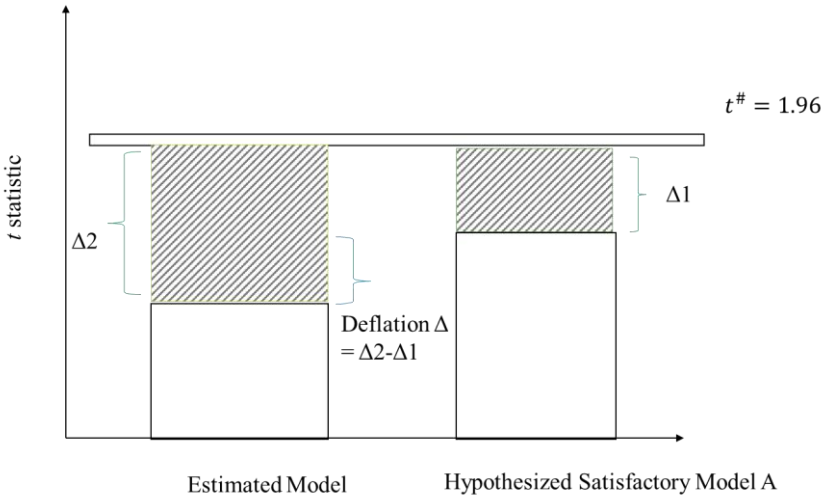
When $t_{ols} < t^{\#} < t_{vif}$, a Type II error occurs, and W_{OC} and ES_{OC} are the same as above. Further, the risk of making a Type II error index is identical to the above Type I error one as:

$$R_{OC} = \frac{\Delta_1}{\Delta} = \frac{\Delta_1}{\Delta_1 + \Delta_2} = \frac{1}{1 + \Delta_2/\Delta_1} = \frac{1}{1 + r},$$

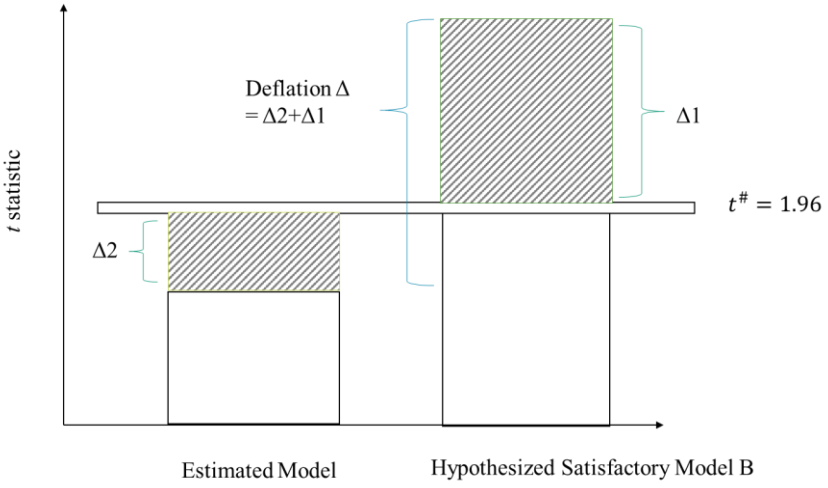
and

$$r = \Delta_2/\Delta_1 = \frac{t^{\#} - t_{ols}}{t_{vif} - t^{\#}} = \frac{StE^{\#} - StE_{ols}\sqrt{VIF}}{(StE_{ols} - StE^{\#})\sqrt{VIF}},$$

where r is positive and $0 < R_{OC} < 1$, since $StE_{ols} > StE^{\#} > StE_{vif}$.



(a) Scenario of no Type II error



(b) Scenario of having Type II error

Figure 3.A.1 Two scenarios of comparing t statistics of the estimated model and the hypothesized satisfactory models ($t_{ols} < t^{\#}$).

APPENDIX 4A

Simulation Parameter Settings and Results of VOCs of Omitting the Highest Cluster Level

Table 4A.1 Simulation parameter settings.

$\widetilde{\boldsymbol{\sigma}}^{(j)}$ **or** $\boldsymbol{\rho} \cong \boldsymbol{\rho_1}$	$\boldsymbol{\sigma}^{(k)}$ **or** $\boldsymbol{\rho_2}$	$\boldsymbol{\sigma}^{(j)}$ **or** $\boldsymbol{\rho_0}$	$\boldsymbol{n_H}$ **Avg. No. of teachers per school**	$\boldsymbol{M_k}$ **No. of schools**	$\widetilde{\boldsymbol{\sigma}}^{(i)} \cong \boldsymbol{\sigma}^{(i)}$
0.5	0.1	0.4	20	5	0.5
0.8	0.4	0.4			0.2
0.8	0.6	0.2			0.2
0.5	0.1	0.4	10	10	0.5
0.8	0.4	0.4			0.2
0.8	0.6	0.2			0.2
0.5	0.1	0.4	4	25	0.5
0.8	0.4	0.4			0.2
0.8	0.6	0.2			0.2
0.5	0.1	0.4	2	50	0.5
0.8	0.4	0.4			0.2
0.8	0.6	0.2			0.2

Table 4A.2 Relative bias of estimates of variances when $\rho_1 = 0.5$, $\rho_2 = 0.1$, $\rho_0 = 0.4$.

Parameter	$\mathbf{n}_H$	M_k	**R. B.$_{est}$ of HLM**		**R. B.$_{adj.est}$ of HLM**	
			Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Teacher-level random effect variance $\tilde{\sigma}^{(j)}$	20	5	0.14 (0.02)	[0, 1.14]	0 (0)	[0, 0]
	10	10	0.19 (0.03)	[0, 1.12]	0 (0)	[0, 0]
	4	25	0.24 (0.03)	[0, 1.06]	0 (0)	[0, 0]
	2	50	0.31 (0.06)	[0, 1.68]	0 (0)	[0, 0]
Standard error of $W_{j(k)}$ coefficient $\tilde{\gamma}_{01}$	20	5	0.06 (0)	[0, 0.43]	0 (0)	[0, 0.02]
	10	10	0.08 (0)	[0, 0.42]	0 (0)	[0, 0.02]
	4	25	0.10 (0.01)	[0, 0.41]	0.01 (0)	[0, 0.03]
	2	50	0.13 (0.01)	[0, 0.58]	0.01 (0)	[0, 0.04]
Standard error of Z_k coefficient $\tilde{\gamma}_{02}$	20	5	-0.33 (0.04)	[-0.69, 0]	0 (0)	[-0.01, 0.01]
	10	10	-0.30 (0.02)	[-0.58, 0]	0 (0)	[-0.01, 0.01]
	4	25	-0.17 (0.01)	[-0.37, 0]	0 (0)	[0, 0]
	2	50	-0.08 (0)	[-0.21, 0.00]	0 (0)	[0, 0]

Table 4A.2 (cont'd)

Parameter	n_H	M_k	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	20	5	0.38 (0)	[0.24, 0.56]	0.02 (0)	[-0.16, 0.07]
	10	10	0.39 (0)	[0.24, 0.56]	0.01 (0)	[-0.15, 0.08]
	4	25	0.40 (0)	[0.27, 0.61]	<0.01 (0)	[-0.22, 0.07]
	2	50	0.40 (0)	[0.25, 0.57]	<0.01 (0)	[-0.18, 0.07]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{01}$	20	5	-0.66 (0)	[-0.71, -0.58]	-0.02 (0)	[-0.11, 0.10]
	10	10	-0.66 (0)	[-0.70 -0.57]	-0.01 (0)	[-0.11, 0.10]
	4	25	-0.66 (0)	[-0.71 -0.55]	<0.01 (0)	[-0.10, 0.11]
	2	50	-0.65 (0)	[-0.71, -0.51]	<0.01 (0)	[-0.10, 0.11]
Standard error of $Z_{i(jk)}$ coefficient $\tilde{\gamma}_{02}$	20	5	-0.79 (0)	[-0.91, -0.64]	-0.03 (0)	[-0.12, 0.10]
	10	10	-0.78 (0)	[-0.87, -0.65]	-0.02 (0)	[-0.11, 0.10]
	4	25	-0.74 (0)	[-0.82, -0.67]	-0.01 (0)	[-0.11, 0.10]
	2	50	-0.71 (0)	[-0.76, -0.65]	<0.01 (0)	[-0.11, 0.11]

Table 4A.3 Relative bias of estimates of variances when $\rho_1 = 0.8$, $\rho_2 = 0.4$, $\rho_0 = 0.4$.

Parameter	n_H	M_k	**R. B.$_{est}$ of HLM**		**R. B.$_{adj.est}$ of HLM**	
			Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Teacher-level random effect variance $\tilde{\sigma}^{(j)}$	20	5	0.59 (0.26)	[0, 3.79]	0 (0)	[0, 0]
	10	10	0.82 (0.20)	[0.00, 2.58]	0 (0)	[0, 0]
	4	25	0.95 (0.17)	[0.15, 3.38]	0 (0)	[0, 0]
	2	50	1.02 (0.23)	[0.00, 3.29]	0 (0)	[0, 0]
Standard error of $W_{j(k)}$ coefficient $\tilde{\gamma}_{01}$	20	5	0.24 (0.03)	[0, 1.16]	0.02 (0)	[0, 0.05]
	10	10	0.33 (0.02)	[0, 0.87]	0.02 (0)	[-0.97, 0.06]
	4	25	0.38 (0.02)	[0.07, 1.07]	0.02 (0)	[0.01, 0.05]
	2	50	0.40 (0.03)	[0.00, 1.04]	0.03 (0)	[0.00, 0.09]
Standard error of Z_k coefficient $\tilde{\gamma}_{02}$	20	5	-0.57 (0.03)	[-0.75, 0]	-0.01 (0)	[-0.02, 0.01]
	10	10	-0.52 (0.01)	[-0.99, 0]	-0.01 (0)	[-0.97, 0.01]
	4	25	-0.35 (0)	[-0.45, -0.15]	0 (0)	[-0.01, 0.01]
	2	50	-0.17 (0)	[-0.25, 0.00]	0 (0)	[0, 0]

Table 4A.3 (cont'd)

Parameter	n_H	M_k	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	20	5	1.02 (0.06)	[0.62, 1.92]	0.21 (0.04)	[-0.68, 0.37]
	10	10	1.12 (0.04)	[0.63, 1.75]	0.13 (0.05)	[-0.91, 0.38]
	4	25	1.19 (0.03)	[0.68, 1.74]	0.05 (0.06)	[-0.97, 0.36]
	2	50	1.20 (0.02)	[0.71, 1.63]	0.03 (0.05)	[-0.83, 0.35]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{01}$	20	5	-0.68 (0)	[-0.74, -0.49]	-0.07 (0.01)	[-0.25, 0.35]
	10	10	-0.66 (0)	[-0.74, -0.54]	-0.03 (0.01)	[-0.22, 0.25]
	4	25	-0.65 (0)	[-0.72, -0.51]	<0.01 (0.01)	[-0.19, 0.33]
	2	50	-0.65 (0)	[-0.74, -0.49]	0.01 (0)	[-0.16, 0.19]
Standard error of $Z_{i(jk)}$ coefficient $\tilde{\gamma}_{02}$	20	5	-0.89 (0)	[-0.94, -0.72]	-0.09 (0.01)	[-0.26, 0.32]
	10	10	-0.88 (0)	[-1.00, -0.73]	-0.05 (0.01)	[-0.24, 0.21]
	4	25	-0.84 (0)	[-0.87, -0.78]	-0.03 (0.01)	[-0.20, 0.28]
	2	50	-0.79 (0)	[-0.82, -0.74]	-0.01 (0)	[-0.18, 0.15]

Table 4A.4 Relative bias of estimates of variances when $\rho_1 = 0.8$, $\rho_2 = 0.6$, $\rho_0 = 0.2$.

Parameter	$\mathbf{n}_H$	M_k	**R. B.$_{est}$ of HLM**		**R. B.$_{adj.est}$ of HLM**	
			Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Teacher-level random effect variance $\tilde{\sigma}^{(j)}$	20	5	1.69 (1.97)	[0.00, 0.39]	<0.01 (0)	[-0.01, 0]
	10	10	2.47 (1.79)	[0.11, 7.11]	<0.01 (0)	[0, 0]
	4	25	2.85 (1.11)	[0.70, 7.44]	<0.01 (0)	[0, 0]
	2	50	3.12 (1.06)	[1.06, 6.77]	<0.01 (0)	[0, 0]
Standard error of $W_{j(k)}$ coefficient $\tilde{\gamma}_{01}$	20	5	0.57 (0.14)	[0.00, 2.28]	0.05 (0)	[0.00, 0.12]
	10	10	0.80 (0.11)	[0.05, 1.78]	0.06 (0)	[0.01, 0.13]
	4	25	0.91 (0.06)	[0.29, 1.85]	0.07 (0)	[0.03, 0.13]
	2	50	0.97 (0.06)	[0.42, 1.69]	0.07 (0)	[0.03 0.17]
Standard error of Z_k coefficient $\tilde{\gamma}_{02}$	20	5	-0.67 (0.01)	[-0.77, 0.00]	-0.02 (0)	[-0.06, 0.01]
	10	10	-0.61 (0)	[-0.66, -0.27]	-0.01 (0)	[-0.04, 0.01]
	4	25	-0.43 (0)	[-0.48, -0.33]	<0.01 (0)	[-0.02, 0.01]
	2	50	-0.24 (0)	[-0.27, -0.18]	<0.01 (0)	[-0.01, 0.00]

Table 4A.4 (cont'd)

Parameter	n_H	M_k	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
			Mean (Variance)	Range	Mean (Variance)	Range
Standard error of $X_{i(jk)}$ coefficient $\tilde{\gamma}_{10}$	20	5	0.91 (0.12)	[0.38, 2.26]	0.23 (0.04)	[-0.89, 0.38]
	10	10	1.07 (0.08)	[0.45, 1.88]	0.18 (0.05)	[-0.87, 0.39]
	4	25	1.18 (0.04)	[0.53, 1.93]	0.09 (0.06)	[-0.93, 0.37]
	2	50	1.20 (0.03)	[0.64, 1.67]	0.04 (0.06)	[-0.89, 0.36]
Standard error of $W_{i(jk)}$ coefficient $\tilde{\gamma}_{01}$	20	5	-0.59 (0.01)	[-0.70, -0.23]	-0.10 (0.02)	[-0.34, 0.57]
	10	10	-0.54 (0)	[-0.69, -0.33]	<0.01 (0.02)	[-0.31, 0.45]
	4	25	-0.52 (0)	[-0.65, -0.33]	0.04 (0.01)	[-0.21, 0.48]
	2	50	-0.51 (0)	[-0.63, -0.33]	0.06 (0.01)	[-0.16, 0.28]
Standard error of $Z_{i(jk)}$ coefficient $\tilde{\gamma}_{02}$	20	5	-0.91 (0)	[-0.94, -0.68]	-0.16 (0.02)	[-0.39, 0.44]
	10	10	-0.90 (0)	[-0.92, -0.78]	-0.07 (0.02)	[-0.33, 0.38]
	4	25	-0.86 (0)	[-0.88, -0.82]	-0.03 (0.01)	[-0.25, 0.38]
	2	50	-0.81 (0)	[-0.82, -0.78]	-0.01 (0)	[-0.22, 0.19]

APPENDIX 5A

Simulation Parameter Settings and Results of VOCs of Omitting the Lowest Cluster Level

Table 5A.1 Simulation parameter settings.

n_t	ρ	$\overline{\rho}$	$\rho_{Time_{ti},AR(1)}$	$\rho_{Time_{ti}}$
6	0.9	0.789	0.5	0.167
	0.7	0.476		
	0.5	0.269		
	0.2	0.079		
10	0.9	0.697	0.7	0.140
	0.7	0.351		
	0.5	0.178		
	0.2	0.049		
30	0.9	0.423	0.9	0.06
	0.7	0.143		
	0.5	0.064		
	0.2	0.017		

Table 5A.2 Relative bias of estimates of variances when lag-1 autocorrelation $\rho = 0.9$

Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of ID**		**R. B.$_{adj.est}$ of ID**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Residual variance $\tilde{\sigma}^{(t)}$	6	-0.79 (0.00)	[-0.81, -0.76]	0.00 (0.00)	[-0.12, 0.12]
	10	-0.70 (0.00)	[-0.73, -0.66]	0.00 (0.00)	[-0.10, 0.11]
	30	-0.42 (0.00)	[-0.47, -0.36]	0.00 (0.00)	[-0.07, 0.10]
Individual level random effect variance $\tilde{\sigma}^{(i)}$	6	1.77 (0.04)	[1.24, 2.23]	0.00 (0.04)	[-0.58, 0.56]
	10	1.57 (0.03)	[1.11,2.09]	0.01 (0.03)	[-0.54, 0.57]
	30	0.95 (0.02)	[0.55, 1.32]	0.00 (0.02)	[-0.38, 0.36]
Standard error of $Time_{ti}$ coefficient $\tilde{\gamma}_{10}$	6	-0.39 (0.00)	[-0.41, -0.36]	-0.04 (0.00)	[-0.12, 0.02]
	10	-0.51 (0.00)	[-0.53, -0.49]	-0.03 (0.00)	[-0.11, -0.04]
	30	-0.71 (0.00)	[-0.72, -0.69]	-0.03 (0.00)	[-0.08, 0.03]
Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of OLS**		**R. B.$_{adj.est}$ of OLS**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Standard error of X_i coefficient $\tilde{\gamma}_{01}$	6	-0.56 (0.00)	[-0.57, -0.55]	0.01 (0.00)	[-0.01, 0.04]
	10	-0.64 (0.00)	[-0.65, -0.63]	0.02 (0.00)	[-0.01, 0.05]
	30	-0.74 (0.00)	[-0.75, -0.73]	0.12 (0.00)	[0.06, 0.17]

Table 5A.3 Relative bias of estimates of variances when lag-1 autocorrelation $\rho = 0.7$

Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of ID**		**R. B.$_{adj.est}$ of ID**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Residual variance $\tilde{\sigma}^{(t)}$	6	-0.48 (0.00)	[-0.53, -0.42]	0.00 (0.00)	[-0.10, 0.10]
	10	-0.35 (0.00)	[-0.41, -0.29]	0.00 (0.00)	[-0.09, 0.10]
	30	-0.14 (0.00)	[-0.19, -0.09]	0.00 (0.00)	[-0.06, 0.06]
Individual level random effect variance $\tilde{\sigma}^{(i)}$	6	1.07 (0.02)	[0.58, 1.53]	0.00 (0.02)	[-0.53, 0.51]
	10	0.79 (0.01)	[0.46, 1.13]	0.00 (0.01)	[-0.30, 0.37]
	30	0.33 (0.01)	[0.09, 0.59]	0.01 (0.01)	[-0.24, 0.28]
Standard error of $Time_{ti}$ coefficient $\tilde{\gamma}_{10}$	6	-0.33 (0.00)	[-0.40, -0.30]	-0.04 (0.00)	[-0.15, 0.01]
	10	-0.41 (0.00)	[-0.44, -0.39]	0.05 (0.00)	[0.01, 0.09]
	30	-0.63 (0.00)	[-0.65, -0.61]	0.01 (0.00)	[-0.04, 0.04]
Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of OLS**		**R. B.$_{adj.est}$ of OLS**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Standard error of X_i coefficient $\tilde{\gamma}_{01}$	6	-0.50 (0.00)	[-0.52, -0.48]	0.02 (0.00)	[-0.01, 0.06]
	10	-0.58 (0.00)	[-0.60, -0.57]	0.02 (0.00)	[-0.01, 0.04]
	30	-0.62 (0.00)	[-0.64, -0.61]	0.36 (0.00)	[0.26, 0.41]

Table 5A.4 Relative bias of estimates of variances when lag-1 autocorrelation $\rho = 0.5$.

Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of ID**		**R. B.$_{adj.est}$ of ID**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Residual variance $\tilde{\sigma}^{(t)}$	6	-0.27 (0.00)	[-0.34, -0.20]	0.00 (0.00)	[-0.09, 0.09]
	10	-0.18 (0.00)	[-0.24, -0.12]	0.00 (0.00)	[-0.08, 0.07]
	30	-0.07 (0.00)	[-0.11, -0.03]	0.00 (0.00)	[-0.05, 0.04]
Individual level random effect variance $\tilde{\sigma}^{(i)}$	6	0.61 (0.01)	[0.28, 0.95]	0.01 (0.01)	[-0.33, 0.36]
	10	0.40 (0.01)	[0.11, 0.67]	0.00 (0.01)	[-0.29, 0.28]
	30	0.15 (0.01)	[-0.07, 0.38]	0.00 (0.01)	[-0.22, 0.23]
Standard error of $Time_{ti}$ coefficient $\tilde{\gamma}_{10}$	6	-0.27 (0.00)	[-0.36, -0.22]	-0.03 (0.00)	[-0.15, 0.05]
	10	-0.32 (0.00)	[-0.35, -0.29]	0.12 (0.00)	[0.07, 0.17]
	30	-0.38 (0.00)	[-0.40, -0.37]	0.58 (0.00)	[0.51, 0.66]
Parameter	$\mathbf{n_t}$	**R. B.$_{est}$ of OLS**		**R. B.$_{adj.est}$ of OLS**	
		Mean (Variance)	**Range**	**Mean (Variance)**	**Range**
Standard error of X_i coefficient $\tilde{\gamma}_{01}$	6	-0.45 (0.00)	[-0.48, -0.39]	0.02 (0.00)	[-0.01, 0.10]
	10	-0.54 (0.00)	[-0.57, -0.52]	0.01 (0.00)	[-0.01, 0.03]
	30	-0.70 (0.00)	[-0.72, -0.68]	0.00 (0.00)	[-0.01, 0.01]

Table 5A.5 Relative bias of estimates of variances when lag-1 autocorrelation $\rho = 0.2$.

Parameter	n_t	R. B.$_{est}$ of ID		R. B.$_{adj.est}$ of ID	
		Mean (Variance)	Range	Mean (Variance)	Range
Residual variance $\tilde{\sigma}^{(t)}$	6	-0.08 (0.00)	[-0.15, 0.01]	0.00 (0.00)	[-0.08, 0.10]
	10	-0.05 (0.00)	[-0.12, 0.02]	0.00 (0.00)	[-0.08, 0.07]
	30	-0.02 (0.00)	[-0.06, 0.02]	0.00 (0.00)	[-0.04, 0.04]
Individual level random effect variance $\tilde{\sigma}^{(i)}$	6	0.17 (0.01)	[-0.11, 0.44]	0.00 (0.01)	[-0.30, 0.27]
	10	0.11 (0.01)	[-0.15, 0.40]	0.00 (0.01)	[-0.26, 0.29]
	30	0.04 (0.00)	[-0.18, 0.24]	0.00 (0.00)	[-0.22, 0.20]
Standard error of $Time_{ti}$ coefficient $\tilde{\gamma}_{10}$	6	-0.12 (0.00)	[-0.16, -0.06]	0.09 (0.00)	[0.02, 0.14]
	10	-0.14 (0.00)	[-0.17, -0.10]	0.29 (0.00)	[0.23, 0.35]
	30	-0.17 (0.00)	[-0.19, -0.15]	1.05 (0.00)	[0.93, 1.12]
Parameter	n_t	R. B.$_{est}$ of OLS		R. B.$_{adj.est}$ of OLS	
		Mean (Variance)	Range	Mean (Variance)	Range
Standard error of X_i coefficient $\tilde{\gamma}_{01}$	6	-0.40 (0.00)	[-0.43, -0.36]	0.00 (0.00)	[-0.01, 0.01]
	10	-0.50 (0.00)	[-0.53, -0.47]	0.00 (0.00)	[-0.01, 0.01]
	30	-0.69 (0.00)	[-0.70, -0.66]	0.00 (0.00)	[0.00, 0.00]

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (No. w24003). National Bureau of Economic Research.

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, *88*(1), 265-296.

Abe, Y., & Gee, K. A. (2014). Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education. *Evaluation and program planning*, *47*, 26-34.

Adelson, J. L., McCoach, D. B., & Gavin, M. K. (2012). Examining the effects of gifted programming in mathematics and reading using the ECLS-K. *Gifted Child Quarterly*, *56*(1), 25-39.

Akerlof, G. A., & Kranton, R. E. (2002). Identity and schooling: Some lessons for the economics of education. *Journal of economic literature*, *40*(4), 1167-1201.

Alejo, J., Montes-Rojas, G., & Sosa-Escudero, W. (2018). Testing for serial correlation in hierarchical linear models. *Journal of Multivariate Analysis*, *165*, 101-116.

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2019). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, 1094428119877457.

Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: University of Chicago Press.

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*(1), 65-74.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*(2), 151.

Bates, D., Maechler, M., Bolker, B., Walker, S., & Haubo Bojesen Christensen, R. (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1–7. 2014.

Battaglia, M. (2008). Multi-stage sample. In P. J. Lavrakas (Ed.). *Encyclopedia of survey research methods* (pp. 493-493). Thousand Oaks, CA: SAGE Publications, Inc. doi: *10.4135/9781412963947.n311*

Baltagi, B. H., & Li, Q. (1991). A joint test for serial correlation and random individual effects. *Statistics & Probability Letters*, *11*(3), 277-280.

Baltagi, B. H., Song, S. H., & Jung, B. C. (2002). Simple LM tests for the unbalanced nested error component regression model. *Econometric Reviews*, *21*(2), 167-187.

Baltagi, B. H., Jung, B. C., & Song, S. H. (2010). Testing for heteroskedasticity and serial correlation in a random effects panel data model. *Journal of Econometrics*, *154*(2), 122-124.

Beaton, A. E., & O'Dwyer, L. M. (2002). Separating school, classroom, and student variances and their relationship to socio-economic status. In D. F. Robitaille & A. E. Beaton, (Eds.). *Secondary analysis of the TIMSS data* (pp. 211-231). Dordrecht: Springer. doi: *10.1007/0-306-47642-8*

Bera, A. K., Sosa-Escudero, W., & Yoon, M. (2001). Tests for the error component model in the presence of local misspecification. *Journal of Econometrics*, *101*(1), 1-23.

Berger, M. P., & Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research* (Vol. 83). John Wiley & Sons.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. *The Quarterly journal of economics*, *119*(1), 249-275.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30-59.

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *The review of economic studies*, *47*(1), 239-253.

Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 159-204). Academic Press.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources*, *50*(2), 317-372.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*(3), 414-427.

Chen, S., & Rust, K. (2017). An extension of Kish's formula for design effects to two-and three-stage designs with stratification. *Journal of Survey Statistics and Methodology*, *5*(2), 111-130.

Cheong, Y. F., Fotiu, R. P., & Raudenbush, S. W. (2001). Efficiency and robustness of alternative estimators for two-and three-level models: The case of NAEP. *Journal of Educational and Behavioral Statistics*, *26*(4), 411-429.

Claessens, A. (2012). Kindergarten child care experiences and child achievement and socioemotional skills. *Early Childhood Research Quarterly*, *27*(3), 365-375.

Coburn, C. E., Russell, J. L., Kaufman, J. H., & Stein, M. K. (2012). Supporting sustainability: Teachers' advice networks and ambitious instructional reform. *American Journal of Education*, *119*(1), 137-182.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Psychology Press.

Conaway, C., Keesler, V., & Schwartz, N. (2015). What research do state education agencies really need? The promise and limitations of state longitudinal data systems. *Educational Evaluation and Policy Analysis*, *37*(1_suppl), 16S-28S.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

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

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009–2015. *Educational Researcher*, *47*(4), 221-234.

Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, *37*(3), 379-403.

Fitchett, P. G., & Heafner, T. L. (2017). Student demographics and teacher characteristics as predictors of elementary-age students' history knowledge: Implications for teacher education and practice. *Teaching and Teacher Education*, *67*, 79-92.

Frank, K. A. (1998). Quantitative methods for studying social context in multilevels and through interpersonal relations. *Review of research in education*, *23*(1), 171-216.

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437-460.

Frank, K. A., Muller, C., Schiller, K. S., Riegle-Crumb, C., Mueller, A. S., Crosnoe, R., & Pearson, J. (2008). The social dynamics of mathematics coursetaking in high school. *American Journal of Sociology*, *113*(6), 1645-1696.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, *141*(1), 2.

Gamoran, A., & Dreeben, R. (1986). Coupling and control in educational organizations. *Administrative science quarterly*, 612-632.

Gamoran, A., Secada, W. G., & Marrett, C. B. (2000). The organizational context of teaching and learning. In *Handbook of the sociology of education* (pp. 37-63). Springer, Boston, MA.

Goddard, Y. L., Goddard, R. D., & Tschannen-Moran, M. (2007). A theoretical and empirical investigation of teacher collaboration for school improvement and student achievement in public elementary schools. *Teachers college record*, *109*(4), 877-896.

Goldstein, H. (2011). *Multilevel statistical models*. Hoboken, NJ: Wiley

Hallinger, P., & Murphy, J. F. (1986). The social context of effective schools. *American journal of education*, *94*(3), 328-355.

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of econometrics*, *140*(2), 670-694.

Heafner, T. L., VanFossen, P. J., & Fitchett, P. G. (2019). Predictors of students′ achievement on NAEP-Economics: A multilevel model. *The Journal of Social Studies Research*, *43*(4), 327-341.

Heck, R. H., Larsen, T. J., & Marcoulides, G. A. (1990). Instructional leadership and school achievement: Validation of a causal model. *Educational Administration Quarterly*, *26*(2), 94-125.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics, 32, 341-370.

Hedges, L. V. (2008). What are effect sizes and why do we need them?. *Child Development Perspectives*, *2*(3), 167-171.

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass Correlations and Covariate Outcome Correlations for Planning Two- and Three-Level Cluster-Randomized Experiments in Education. *Evaluation Review*, *37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Hedges, L. V., & Rhoads, C. (2010). Statistical Power Analysis in Education Research. NCSER 2010-3006. *National Center for Special Education Research*.

Heo, M., & Leon, A. C. (2008). Statistical power and sample size requirements for three level hierarchical cluster randomized trials. *Biometrics*, *64*(4), 1256-1262.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. American educational research journal, 42(2), 371-406.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child development perspectives*, *2*(3), 172-177.

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012, July). How few countries will do? Comparative survey analysis from a Bayesian perspective. In *Survey Research Methods* (Vol. 6, No. 2, pp. 87-93).

Huang, F. L. (2018). Multilevel modeling myths. *School Psychology Quarterly*, *33*(3), 492.

Ilies, R., & Judge, T. A. (2004). An experience-sampling measure of job satisfaction and its relationships with affectivity, mood at work, job beliefs, and general job satisfaction. *European journal of work and organizational psychology*, *13*(3), 367-389.

Jayanthi, M., Dimino, J., Gersten, R., Taylor, M. J., Haymond, K., Smolkowski, K., & Newman-Gonchar, R. (2018). The impact of teacher study groups in vocabulary on teaching practice, teacher knowledge, and student vocabulary knowledge: A large-scale replication study. *Journal of Research on Educational Effectiveness*, 11(1), 83-108.

Jennings, J. L., & DiPrete, T. A. (2010). Teacher effects on social and behavioral skills in early elementary school. *Sociology of Education*, *83*(2), 135-159.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 983-997.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583-2595.

King, G., & Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 159-179.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics, 11*(1), 55. Retrieved from http://ezproxy.msu.edu.proxy1.cl.msu.edu/login?url=https://search-proquest-com.proxy1.cl.msu.edu/docview/1266820489?accountid=12598

Konstantopoulos, S. (2008a). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*(1), 66-88.

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*(4), 265-288.

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster randomized designs. *Evaluation Review*, *33*(4), 335–57. https://doi.org/10.1177/0193841X09337991

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education*, *78*(3), 291-317.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61-76.

Korendijk, E. J., Hox, J. J., Moerbeek, M., & Maas, C. J. (2011). Robustness of parameter and standard error estimates against ignoring a contextual effect of a subject-level covariate in cluster-randomized trials. *Behavior research methods*, *43*(4), 1003-1013.

Korendijk, E. J., Moerbeek, M., & Maas, C. J. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics*, *35*(5), 566-585.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241-253.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research*, *36*(2), 249-277.

Kwok, O. M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*(3), 557-592.

LeBeau, B. (2016). Impact of serial correlation misspecification with the linear mixed model. *Journal of Modern Applied Statistical Methods*, *15*(1), 21.

LeBeau, B. (2018, May). *Misspecification of the random effect structure: Implications for the linear mixed model [Working paper]. Retrieved January 20, 2020 from* https://iro.uiowa.edu/discovery/fulldisplay/alma9983557687702771/01IOWA_INST:ResearchRepository?tags=scholar

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. doi: *10.3102/1076998614546494*

Leeuw J., & Meijer E. (2008) Introduction to Multilevel Analysis. In J. De Leeuw & E. Meijer (Eds.). *Handbook of Multilevel Analysis* (pp. 1-75). New York, NY: Springer.

Lendrum, A., & Humphrey, N. (2012). The importance of studying the implementation of interventions in school settings. *Oxford Review of Education*, *38*(5), 635-652.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, *19*(13), 1793-1819.

Luo, W., & Kwok, O. M. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*(2), 182-212.

MacKinnon, J. G., & Webb, M. D. (2019, May). *When and how to deal with clustered errors in regression models* (Queen's Economics Department Working Paper No. 1421). *Retrieved January 20, from http://qed.econ.queensu.ca/pub/faculty/mackinnon/working-papers/qed_wp_1421.pdf*

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: an illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement*, *23*(3), 305-326.

McNeish, D. M. (2014). Analyzing clustered data with OLS regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, *40*(1), 11-16.

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.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114.

McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research*, *52*(2), 200-215.

Moeller, A. J., Theiler, J. M., & Wu, C. (2012). Goal setting and student achievement: A longitudinal study. *The Modern Language Journal*, *96*(2), 153-169.

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate behavioral research*, *39*(1), 129-149.

Montes-Rojas, G. (2016). An equicorrelation Moulton factor in the presence of arbitrary intra-cluster correlation. *Economics Letters*, *145*, 221-224.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of econometrics*, *32*(3), 385-397.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *The review of Economics and Statistics*, 334-338.

Muijs D. (2020) Extending Educational Effectiveness: The Middle Tier and Network Effectiveness. In J. Hall, A. Lindorff, & P. Sammons (Eds.). *International Perspectives in Educational Effectiveness Research*. Springer, Cham. doi: *10.1007/978-3-030-44810-3_5*

Muller, C. L. (2015). Measuring school contexts. *AERA open*, *1*(4). https://doi.org/10.1177/2332858415613055

Mundfrom, D. J., & Schults, M. R. (2002). A Monte Carlo simulation comparing parameter estimates from multiple linear regression and hierarchical linear modeling. *Multiple Regression Viewpoints*, *28*, 18-21.

Murphy, D. L., & Pituch, K. A. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*, *77*(3), 255-284.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cornfield's penalties?. Evaluation Review, 20(3), 313-337.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., & Brauer, M. (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology*, *74*(2).

Nezlek, J. B., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze grouped data. *Group Dynamics: Theory, Research, and Practice*, *2*(4), 313.

Niehaus, E., Campbell, C. M., & Inkelas, K. K. (2014). HLM behind the curtain: Unveiling decisions behind the use and interpretation of HLM in higher education research. *Research in Higher Education*, *55*(1), 101-122.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. Educational evaluation and policy analysis, *26*(3), 237-257.

Opdenakker, M. C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, *11*(1), 103-130.

OECD (2017). *PISA 2015 technical report*. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf

Penuel, W. R., Riel, M., Krause, A., & Frank, K. A. (2009). Analyzing teachers' professional interactions in a school as social capital: A social network approach. *Teachers college record*, *111*(1), 124-163.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206-230.

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

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*(2), 138-154.

Raudenbush, S. W., & Schwartz, D. (2020). Randomized Experiments in Education, with Implications for Multilevel Causal Inference. *Annual Review of Statistics and Its Application*, 7, 177-208.

Raykov, T., Patelis, T., Marcoulides, G. A., & Lee, C. L. (2016). Examining intermediate omitted levels in hierarchical designs via latent variable modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 111-115.

Rhoads, C. H. (2011). The implications of "contamination" for experimental design in education. *Journal of Educational and Behavioral Statistics*, *36*(1), 76-104.

Robinson, T. Three essays on measuring political behaviour [PhD thesis]. University of Oxford.

Rumberger, R. & Palardy, G (2004). Multilevel models for school effectiveness research. In D. Kaplan (Ed). *The Sage handbook of quantitative methodology for the social sciences* (pp. 235-257). Thousand Oaks, CA: Sage.

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs.

Seashore Louis, K., & Lee, M. (2016). Teachers' capacity for organizational learning: The effects of school culture and context. *School Effectiveness and School Improvement*, *27*(4), 534-556.

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

Skinner, C. J. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, *48*(1), 89-99.

Skinner, C. J., Holt, D., & Smith, T. F. (1989). *Analysis of complex surveys*. John Wiley & Sons.

Skinner, C., & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165-175.

Skrondal A., & Rabe-Hesketh S. (2008) Multilevel and Related Models for Longitudinal Data. In J. De Leeuw & E. Meijer (Eds.). *Handbook of Multilevel Analysis* (pp. 275-300). New York, NY: Springer.

Snijders T.A., & Berkhof, J. (2008) Diagnostic Checks for Multilevel Models. In J. De Leeuw & E. Meijer (Eds.). *Handbook of Multilevel Analysis* (pp. 141-175). New York, NY: Springer.

Snijders T.A. (2005). Power and sample size in multilevel modeling. In B.S. Everitt & D.C. Howell (Eds.). *Encyclopedia of Statistics in Behavioral Science* (pp. 1570–1573*)*. Vol. 3. Chichester, UK: Wiley. doi: *10.1002/0470013192*

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Spillane, J. P., Parise, L. M., & Sherer, J. Z. (2011). Organizational routines as coupling mechanisms: Policy, school administration, and the technical core. *American educational research journal*, *48*(3), 586-619.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605-627.

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from cluster randomized trials in education: An assessment of the capacity of studies to determine what works, for whom, and under what conditions. *Educational Evaluation and Policy Analysis*, *42*(3), 354-374.

Stapleton, L. M., & Kang, Y. (2018). Design effects of multilevel estimates from national probability samples. *Sociological Methods & Research*, *47*(3), 430-457.

Taylor, I. M., Ntoumanis, N., Standage, M., & Spray, C. M. (2010). Motivational predictors of physical education students' effort, exercise intentions, and leisure-time physical activity: A multilevel linear growth analysis. *Journal of Sport and Exercise Psychology*, *32*(1), 99-120.

Tranmer, M., & Steel, D. G. (2001). Ignoring a level in a multilevel model: evidence from UK census data. *Environment and Planning A*, *33*(5), 941-948.

Vallejo, G., Fernández, M. P., Livacic-Rojas, P. E., & Tuero-Herrero, E. (2011). Selecting the best unbalanced repeated measures model. *Behavior research methods*, *43*(1), 18-36.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). Practical tools for designing and weighting survey samples. New York: Springer.

van Breukelen, G & Moerbeek, M. (2016). Design considerations in multilevel studies. In M. A. Scott, J. S. Simonoff & D. Marx (Eds.). *The SAGE handbook of Multilevel Modeling* (pp. 183-200). London, UK: SAGE Publications Ltd. doi: *10.4135/9781446247600*

Van den Noortgate, W., Opdenakker, M. C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*(3), 281-303.

Vaezghasemi, M., Ng, N., Eriksson, M., & Subramanian, S. V. (2016). Households, the omitted level in contextual analysis: disentangling the relative influence of households and districts on the variation of BMI about two decades in Indonesia. *International journal for equity in health*, *15*(1), 102.

Voogt, J. M., Pieters, J. M., & Handelzalts, A. (2016). Teacher collaboration in curriculum design teams: effects, mechanisms, and conditions. *Educational Research and Evaluation*, *22*(3-4), 121-140.

Wang, W., Liao, M., & Stapleton, L. (2019). Incidental Second-Level Dependence in Educational Survey Data with a Nested Data Structure. *Educational Psychology Review*, 1-26.

Weiss, M. J. (2010). The implications of teacher selection and the teacher effect in individually randomized group treatment trials. *Journal of Research on Educational Effectiveness*, *3*(4), 381-405.

Weiss, M. J., Lockwood, J. R., & McCaffrey, D. F. (2016). Estimating the standard error of the impact estimator in individually randomized trials with clustering. *Journal of Research on Educational Effectiveness*, *9*(3), 421-444.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. https://doi.org/10.1177/0193841X14531584

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817-838.

Wong, E. M., & Li, S. C. (2008). Framing ICT implementation in a context of educational change: A multilevel analysis. *School effectiveness and school improvement*, *19*(1), 99-120.

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review*, *93*(2), 133-138.

Xia, J., Shen, J., & Sun, J. (2020). Tight, loose, or decoupling? A National study of the decision-making power relationship between district central offices and school principals. *Educational Administration Quarterly*, *56*(3), 396-434.

Zhu, P., Jacob, R., Bloom, H., & Xu, Z. (2012). Designing and analyzing studies that randomize schools to estimate intervention effects on student academic outcomes without classroom-level information. *Educational Evaluation and Policy Analysis*, *34*(1), 45-68.