QUANTIFYING STRENGTH OF EVIDENCE IN EDUCATION RESEARCH: ACCOUNTING FOR SPILLOVER, HETEROGENEITY, AND MEDIATION

By

Qinyun Lin

A DISSERTATION

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

Measurement and Quantitative Methods – Doctor of Philosophy

2019

ABSTRACT

QUANTIFYING STRENGTH OF EVIDENCE IN EDUCATION RESEARCH: ACCOUNTING FOR SPILLOVER, HETEROGENEITY, AND MEDIATION

By

Qinyun Lin

It is very rare that education studies have constant intervention effects through simple mechanisms to independent individuals. It is well-documented that schooling is a complex process because teachers, students, and administrators interact with each other in a diverse set of social contexts (e.g., An, 2018; Frank, 1998; Hong, 2015; Kim, Frank, & Spillane, 2018; Maroulis et al., 2010). As such, considering potential bias due to unobserved or uncontrolled spillover, heterogeneity and alternative mediators is important to making an inference for policy implications. Additionally, since the ultimate goal of education research is to inform decision-makings in the allocation of educational resources regarding curricula, pedagogy, practices or school organizations (e.g., Bulterman-Bos, 2008; Cook, 2002), education research must be accessible to practitioners. Consequently, a sensitivity framework that can account for all potential sources of bias, including spillover, heterogeneity and alternative mediators, is required to allow all stakeholders to conceptualize the quality of evidence independently so that the debate for future policy manipulations can take place in a more transparent, effective and equitable way.

Drawn on the work by Frank, Maroulis, Duong, and Kelcey (2013), Chapters 1 and 2 in this dissertation propose a non-parametric case replacement approach to quantify the robustness of inference in multisite randomized control trials and value-added measures for teacher effectiveness, accounting for spillover and heterogeneity. Throughout, the Tennessee class size experiment (Project STAR) is applied to demonstrate the case replacement approach. Chapters 3 and 4 focus on unobserved mediators in a single-mediator model. Specifically, Chapter 3 examines whether and how omitting an alternative mediator can bias causal mediation effect estimates in a cross-sectional single-mediator model. Further, a sensitivity analysis approach is proposed to evaluate the robustness of causal mediation inference to missing a potential confounding mediator. Chapter

4 continues the discussion in Chapter 3 and a parameter framework is developed to characterize inconsistency in mediation models. This parameter framework is also applied to a longitudinal design for a post-treatment confounder.

This dissertation is dedicated in memory of my mother, Minhong Du. I miss you everyday, but I believe you would be glad to see this process through to completion.

ACKNOWLEDGMENTS

Doctoral study is such a long and challenging journey that I would have not been here without help and support from so many great people. I have been waiting for this moment when I can say THANK YOU to all the great people that I have met and learned from during the past five years.

First, I would like to show my greatest gratitude to my incredible adviser, Dr. Ken Frank, for all his encouragement, guidance, trust, and support. I have learned so much from him that is far beyond just doing good research. I was so fortunate to have so many opportunities to travel with him and talk with him to learn about his journey as a scholar, a professor and a farther, which has inspired me greatly along my journey as a graduate student, especially when I felt struggled or lack of confidence for my future as a scholar. He has brought me to the world where I found genuine joy in research, especially those moments when I finally figured out those intriguing questions. Thank you so much for always believing and trusting me. It is your encouragement and support that has given me the confidence to continue pursuing the research journey I have dreamed about!

I want to express my sincere appreciation to all my committee members, for all the constructive suggestions and advice that they have provided. My dissertation is very interdisciplinary as it relates to methods and topics across education, econometrics and psychology. It is my committee members' open-mindedness that has made this dissertation possible. I would like to thank Dr. Amy Nuttall for her dedicated and motivating guidance, that has not only made her a great instructor, but also an inspiring mentor that led me through a challenging period of graduate studies. For the past two years, she has been so generous with her time and excellent advice. I also want to express my appreciation for Dr. Jeffrey Wooldridge. Your econometrics courses have introduced me to such an amazing world that I got fascinated by the beauty of quantitative methods. I believe these are the best courses I have ever taken! I also owe a thank you to Dr. Spiro Maroulis and Dr. Spyros Konstantopoulos. They have provided me with so many constructive suggestions and comments on how to frame this dissertation to keep moving forward. Although Dr. Qian Zhang is not my committee member, she has provided me with a great amount of comments and suggestions for the last two chapters, which I appreciate a lot!

I also want to send special thanks to Dr. Andy Anderson for your generous support. You have been a role model for me as a good scientist and researcher. Every time we discussed quantitative analysis, your questions pushed me to think more and understand deeper. I also feel so fortunate to have the opportunity to work with Christie Thomas, Dr. Stefanie Marshall, and everyone else in Carbon TIME in the past five years. I have learned so much from working with you. Thank you for all the help and support!

I also owe so many thanks to my friends and colleagues. Dr. Siwen Guo and Dr. Ran Xu, I appreciate all your help along the way. Talking with you can always help me understand a problem much better when I got stuck at somewhere. I also have special thanks to everyone in our research group: Tingqiao Chen, Zixi Chen, Yuqing Liu, Dr. I-Chien Chen. I deeply appreciate so many discussions and emotional support that we have shared.

Finally, I would like to give my most special thanks to my husband, Xukun Xiang, and my farther, Hui Lin. You are always there with me, no matter when I am struggled, lost, or enthusiastic about some new progress. I am the luckiest person in the world to have such a great family in my life, always supporting me to pursue the life I really want.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This dissertation is centered on causal inference regarding evaluation of an intervention, from whether an intervention works to why it works, accounting for the social and dynamic contexts in which interventions are implemented. The first two chapters propose an approach to quantify the robustness of inference in multisite randomized control trials and value-added measures for teacher effectiveness. Drawn on the work by Frank et al. (2013), these two chapters further extend the non-parametric case replacement approach to quantify how much bias due to spillover violations of SUTVA and presence of heterogeneous treatment effects must be present to invalidate an inference. Throughout, the Tennessee class size experiment (Project STAR) is applied to demonstrate the case replacement approach in both contexts of multisite randomized control trials and value-added measures.

One of the main goals of Project STAR is to study the effect of small class size effect on student achievement. It is well documented that small class size has a positive effect on boosting students' learning, but the effect can vary substantially depending on grade levels and schools (e.g., Hanushek, 1999; Konstantopoulos, 2011; Pedder, 2006; Schanzenbach, 2007). As shown in Chapter 1, this also includes the fact that some schools in Project STAR show strong evidence for negative small class size effects. Additionally, we need to evaluate each school on its own, and to do so we need to quantify its effect relative to a unique threshold for that school. We also need to account for how students might influence each other within or between classrooms as these may bias the estimation of the intervention effects.

Mediating Mechanisms

Underlying this treatment effect heterogeneity are potential mechanisms for how an intervention affects students' learning. Specifically, these mechanisms require in-depth studies of complex classroom and school processes that mediate interventions on students' learning (Pedder, 2006), where various factors are included regarding teachers' practices, classroom discourse routines, teacher-student interactions, peer relations etc. Studying these mechanisms can tell us why an intervention works or not in certain contexts so that future policy manipulations can be better informed. For example, Harfitt (2013) shows that reducing small class sizes may not work if teachers do not seek to exploit the advantages of a smaller class size via changing their pedagogies. With teaching practices serving as a crucial mediator, class size reduction may only work as expected when coupled with professional development for teachers. As such, studying the causal mechanism may allow us to explain the heterogeneity of treatment effects, and to figure out necessary conditions for realizations of intervention effects, so that later policy manipulations can be better informed. Otherwise, heterogeneous treatment effects may lead to unexplained inconsistencies that allow politicians of different persuasions to intentionally select findings that support their preferred policy choices (Blatchford & Martin, 1998; Pedder, 2006).

Recognizing the importance of studying mediation analysis, Chapters 3 and 4 focus on unobserved mediators in a single-mediator model. Specifically, Chapter 3 examines whether and how omitting an alternative mediator that is confounded with an observed mediator can bias causal mediation effect estimates in a cross-sectional single-mediator model. Further, a sensitivity analysis approach is proposed to evaluate the robustness of causal mediation inference to missing a potential confounding mediator. Chapter 4 continues the discussion in Chapter 3 about an unobserved mediator but further leverages a parameter framework to discuss how the bias (more precisely, inconsistency) is generated for each path coefficient of interest in a time varying model consistent with a dynamic process. Applying the Law of Iterated Expectation and the Linear Regression Framework, the bias (more precisely, inconsistency) generation mechanism underlying the cross-sectional model can also be applied to a post-treatment confounder in a time varying single-mediator model.

CHAPTER 1

QUANTIFYING STRENGTH OF EVIDENCE FOR INFERENCES IN MULTISITE RANDOMIZED CONTROL TRIALS: CASE REPLACEMENT, SPILLOVER, AND HETEROGENEITY

1.1 Introduction

In the introduction I described how threats to validity can be interpreted in terms of the sampling mechanism. The goal of this chapter is to characterize the robustness of a causal inference by interpreting it as the percentage of a sample that must be replaced with counterfactual no-effect cases to alter the inference. Most importantly, I will extend this case replacement framework to attend to violations of the Stable Unit Treatment Value Assumption (SUTVA) and presence of heterogeneous treatment effects.

The ultimate goal of any educational research is to inform decision-makings in the allocation of educational resources regarding curricula, pedagogy, practices or school organizations (e.g., Bulterman-Bos, 2008; Cook, 2002). Consequently, education research must be accessible to practitioners. There are two overarching principles underlying the AERA's "Standards for Reporting on Empirical Social Science Research" guide education researchers to engage stakeholders: the sufficiency of the warrants and the transparency of the report. But it is not easy for education researchers to achieve these two principles in practice, especially for varied audiences that may include stakeholders from various backgrounds: policymakers and practitioners including administrators, teachers and parents. Effective communication requires a framework to inform discussions and debate about inferences that make sense to all stakeholders.

Sensitivity analyses can serve as a useful tool to inform debate about specific inferences by quantifying the strength of evidence in education research. The quality of evidence is quantified by discussing the conditions that would alter the inference (e.g., Frank, 2000; Imbens, 2003; Rosenbaum, 2002; VanderWeele & Arah, 2011). These analyses generate statements such as "an omitted variable would have to be correlated at __ with the treatment and with the outcome to invalidate an inference of an effect of the treatment on the outcome." As such recent approaches to sensitivity analysis help interpreters of research quantify the conditions necessary to invalidate an inference drawing on familiar quantities such as correlations (Frank, 2000), percentage of variance explained (Cinelli & Hazlett, 2018) or graphical representations such as contour plots (Imbens, 2003).

But these existing sensitivity analyses approaches are constrained by specific models and the discourse is in the language of correlations and variances. We argue that a well-designed sensitivity analysis framework should go beyond the constraint of specific models and make sense to varied audiences, including those without any statistics background. As a result, a powerful sensitivity framework should allow all stakeholders to conceptualize the quality of evidence independently so that the debate can take place in a more transparent, effective and equitable way. Accordingly, the resulting policy manipulations can also be based on a well-informed discussion. In addition, the advantage of going beyond model constraints allows comparisons among different studies.

As an attempt to provide a powerful sensitivity analysis tool that serves these requirements, we will introduce a non-parametric case replacement approach illustrated by Frank et al. (2013) that draws on Rubin's causal model (RCM) (Rubin, 1974) to express concerns about bias in terms of the characteristics of unobserved, counterfactual data. To demonstrate how this case replacement framework differs from other existing approaches, we will contextualize our discussions in multisite randomized control trials (MSTs). The MSTs are highly relevant as they provide evidence for inference to inform policy manipulations and demonstrate typical education scenarios where SUTVA and constant treatment effect assumption are rarely satisfied in practice.

For purpose of illustration, we will use the Tennessee class size experiment, or Project STAR (Student-Teacher Achievement Ratio), to demonstrate our sensitivity approach. There were 79 elementary schools in 42 school districts involved in this 4-year long project to study the effect of class sizes on student achievement. In each school, kindergarten students were randomly assigned into small classes (13-17 students), regular classes (22-26 students), or regular classes with a full-time aid. Teachers were also randomly assigned to these different types of classes. The assignments of students and teachers to different class types were maintained from kindergarten through the third grade.

1.2 Multisite Randomized Control Trials (MSTs)

Regarded as the “gold standard” and the most powerful experimental design, randomized control trials (MSTs) are being applied with increasing frequency to measure the effectiveness of educational interventions. More than 160 evaluations that randomized individuals or groups to treatment and control conditions have been funded by the National Center for Education Research of the Institute of Education Sciences (IES) since 2002 (Bloom & Spybrook, 2017). Among these, MSTs are gaining increasing popularity (Spybrook & Raudenbush, 2009; Spybrook, Shi, & Kelcey, 2016), where individuals within each site are randomly assigned to treatment and control conditions. With a large and diverse sample from different sites, MSTs may have several potential advantages, including stronger generalizability of findings and allowing estimations of cross-site treatment effect variation (e.g., Bloom, Raudenbush, Weiss, & Porter, 2017; Bloom & Spybrook, 2017). The findings regarding the overall mean treatment effect as well as the variance of cross-site treatment effects are then both used to inform later policy manipulations. For example, some studies based on Project STAR suggest that the positive effect of small class size on student achievement in early grade is large enough to inform education policy (Nye, Hedges, & Konstantopoulos, 2000), but the small class effect is not consistent in all schools: although students in many schools benefit considerably, in other schools being assigned in small class shows no effect or even negative effects on student achievement (Konstantopoulos, 2011). Thus, based on these findings, a policymaker might conclude that reduction of class sizes might benefit some students, but not all.

Even randomized control trials are not free of bias. The underlying idea of randomization is to eliminate possible contaminating effects by trying to ensure no systematic differences in participants’ baseline characteristics. But randomization cannot exclude other sources of error that can happen in educational settings, such as non-compliance, attrition and problems in intervention implementation fidelity (e.g., Hanushek, 1999; Sullivan, 2011). These potential sources of bias can create validity problem of MSTs as well. For example, it is well-documented that Project STAR suffers from a number of important design and implementation issues that can create potential bias (Hanushek, 1999; Konstantopoulos, 2011; Nye et al., 2000). First, the manipulation of class size, as the intervention of interest, was not implemented with fidelity in all schools. Second, there was sizable attrition as well as missing test scores in each year. When attrition occurred, new students were added but there were no pretests available for these new students to verify the randomization through the experiment. Third, the participating schools were not randomly selected, instead they had to volunteer to participate. Evidence shows the sample does differ from the total student population in Tennessee in fall 1986 (Hanushek, 1999). Finally, some students moved between different classroom types (treatment or control conditions in this project) throughout the experiment.

1.3 Spillover and Heterogeneity in MSTs

MSTs demonstrate typical education scenarios where SUTVA is rarely satisfied in practice. As a fundamental assumption for causal inferences, SUTVA requires the treatment status experienced by one unit does not affect the treatment effect for another (Rubin, 1986, 1990). But it is well-documented that schooling is a complex process because teachers, students, and administrators interact with each other in a diverse set of social contexts (e.g., An, 2018; Frank, 1998; Kim et al., 2018; Maroulis et al., 2010). In many MSTs for educational interventions, each site can be one classroom or one school, where peer effects can create bias for even the estimation of the treatment in a single site. For example, in Project STAR, students were randomly assigned to each treatment condition in kindergarten but because of attrition, new students were added to each grade every year. With no guarantee that these new students were randomly assigned, they might become distractors or contributors that can affect other students in the same classroom. It is also possible that students from different classrooms (i.e., treatment conditions) might interact and learn from each other, introducing bias to estimating the between-class achievement average (i.e., intervention effect).

Whenever a single treatment effect is estimated, there can be heterogeneous treatment effects. The heterogeneity can stem from differences in any aspect that relates to the realization of the treatment effect, including individual characteristics, contextual effects and mediating mechanisms. In MSTs, heterogeneous treatment effects can appear both within sites and across sites, among which the cross-site inconsistencies of treatment effects can play a crucial role in the generalizability of the findings and policy implications.

1.4 Strength of Evidence in MSTs

In MSTs, the estimated effects are compared to a certain threshold to make an inference as a basis for policy implications. Specifically, the thresholds based on statistical significance can be applied to claim whether the overall mean intervention effect is significantly positive/negative and whether there are significant cross-site variations. Regardless of the specific definition, a threshold represents "the point at which evidence from a study would make one indifferent to the policy choices" Frank et al. (2013). As argued in (Frank et al., 2013), the comparison between the threshold and the estimate then represents the strength of evidence that supports the inference that directly links to the policy choice. Thus, all stakeholders should be able to understand this comparison so that the policy choice can be made after comprehensive consideration and evaluation of the strength of evidence against potential costs.

For example, if Project STAR supports an inference for a positive small class size effect on student achievement, then future educational policies would be informed to reduce class sizes. However, there are debates about whether the inference is strong enough to inform policies. For instance, Hanushek (1999) maintains that considerable uncertainty about the class size effects is suggested by a number of important design and implementation issues in the project and the evidence is not strong enough to show a systematic effect from overall class size reduction policies. In contrast, Nye et al. (2000) argues that even with shortcomings in implementation, the estimated class size effects are large enough to inform policies. The debate here is essentially about how far the estimated effect exceeds the threshold, and how consistent the effect is across different sites.

Thresholds based on statistical significance require thinking about a repeated sampling framework that conjures a scenario that is beyond the observed data. This can create difficulties for people without any statistical background. As we will demonstrate, the case replacement approach proposed in Frank et al. (2013) provides a more intuitive alternative to quantify the comparison between the threshold and the estimated effect to inform discussions among all stakeholders. Additionally, this framework can be applied to any type of threshold.

This study has two goals. First, we want to demonstrate how the case replacement approach can be applied in MSTs. Second, we aim to extend Frank et al. (2013) work by discussing how we can quantify the uncertainty to violations of SUTVA and presence of heterogeneous treatment effects. In the following section, we will first review the case replacement framework proposed by Frank et al. (2013).

1.5 Case Replacement as a Counterfactual Thought Experiment

In Frank et al. (2013), the authors showed an approach to quantify how much bias there must be in an estimate to invalidate an inference. Then the bias is interpreted in terms of sample replacement to inform more intuitive interpretation. In other words, to show how robust an inference is, we ask a question based on a thought experiment which is counterfactual: what percentage of the sample should be replaced with counterfactual (unobserved) no-effect cases to invalidate an inference made from the data? Or if the concern is about external validity, we consider what percentage of the samples should be replaced with no-effect cases from an unsampled population. The larger the percentage is, the more robust the conclusion/inference is, the less likely that the finding is only due to chance or bias.

This case replacement idea can be applied in various ways to characterize the strength of evidence in MSTs. But the general idea is always about replacing some observed cases with some unobserved cases. The replacement process can become very flexible depending on: (1) how we select cases from the observed sample to be replaced (2) what cases are regarded as replacement cases (3) whether the non-replaced cases in the sample experience any changes during the replacement. In the following analysis, we will discuss different ways to consider the hypothetical replacement and how each approach helps inform the comparison between the threshold and the estimated effect under different contexts.

1.6 Case Replacement for Quantifying the Strength of Evidence in MSTs

1.6.1 Sources of bias in MSTs.

As discussed above, there can be various sources of bias in MSTs. However, we can regard all these as essentially violating the random assignment assumption. That is, sources of bias create differences between the treatment and control group in addition to the experiment condition and more importantly, these differences affect the outcome measures. For example, attrition and added students in Project STAR may introduce differences between the small and regular classes that can lead to differential achievement outcomes. Teacher expectation or reactions to the treatment condition can be another example for one potential contaminating factor. When these differences between treatment and control groups are present, they are confounded with the intervention of interest and we cannot tell whether and how the intervention of interest causes changes separately from other contaminating factors.

1.6.2 Case replacement for each site when SUTVA holds.

In a counterfactual framework, the violation of random assumption indicates the control group individuals do not provide an accurate approximation of counterfactuals for treatment group individuals if they were assigned to the control condition. Or vice versa: the treatment group individuals do not provide accurate approximation of counterfactuals for control group individuals if they had been exposed to the treatment condition. In other words, bias is introduced because 1) treatment group members are compared with observed control group members rather than their counterfactuals: treatment group members if they had received the control and 2) control group members are compared with observed treatment group members rather than their counterfactuals: control group members if they had received the treatment (Frank et al., 2013). That is, the way different sources of biases create problems can be understood as missing data for individuals if they were assigned to a different experiment condition (Holland, 1986). This feature that recasts all potential sources of bias in terms of missing data has been used in the case replacement approach to quantify the robustness of inference (Frank et al., 2013).

To illustrate, consider in each site we have two experiment conditions: treatment vs control. The null hypothesis is $H_0 : \mu_T = \mu_C$, where μ_T and μ_C are population means under the treatment and control conditions, respectively. Note these are different outcomes for the same population under different conditions. Now consider we observe sample averages of treatment and control groups, denoted as $\bar{X}_T$ and $\bar{X}_C$, respectively, and we have $\bar{X}_T - \bar{X}_C > Threshold$ for a significantly positive intervention effect. In other words, $\bar{X}_C$ is used as an approximation of counterfactuals for treatment individuals' outcomes if they had received the control. Now in the thought experiment, assume a proportion of observations, say π, in the control group that cannot serve as accurate approximations for counterfactuals of the treatment group, due to violations of the random assignment assumption. And the true intervention effect for these individuals are zero, indicating their true counterfactuals would be $\bar{X}_T$ rather than $\bar{X}_C$ if they had received the control. Then the average for the control condition becomes: $\bar{X}_C \cdot (1 - \pi) + \bar{X}_T \cdot \pi$. The intervention effect, as the difference between the control and treatment conditions, becomes $\bar{X}_T - [\bar{X}_C \cdot (1 - \pi) + \bar{X}_T \cdot \pi]$. Setting this difference to be equivalent to the threshold so that the amount of bias can invalidate the inference, we can solve for π, where $\pi = 1 - Threshold/(\bar{X}_T - \bar{X}_C)$. That is, π characterizes the amount of bias necessary to invalidate the inference (i.e., the difference between the estimated effect and the threshold), as the proportion of treatment cases for which the true treatment effect is zero and the control group members provide biased approximations of their counterfactuals if they had been assigned to the control condition.

To put it more simply, the case replacement approach quantifies the robustness of an inference by considering replacing a proportion of the observed control cases with unobserved counterfactuals of the treatment cases if they had received the control, assuming these treatment cases experience null treatment effects. This leads to a new estimated effect after replacement: $(\bar{X}_T - \bar{X}_C) \cdot (1 - \pi) + 0 \cdot \pi$.

By setting this to be equivalent to the threshold, we can solve for π that represents a value that can be applied to characterize the strength of evidence against a certain threshold for an inference. A larger π indicates more cases need to be replaced with null-effect cases to invalidate the inference, and correspondingly, more bias needs to be present to invalidate the inference, which represents a more robust inference for all sources of bias. This is a very brief review for how the case replacement approach applies the counterfactual framework to quantify the robustness of inference. See Frank et al. (2013) for more detailed formalization of this thought experiment.

Similar arguments can be applied for scenarios where the observed treatment effect is positive but below the threshold: $0 < \bar{X}_T - \bar{X}_C < Threshold$. Now the goal is to quantify how the data must change to sustain an inference. Specifically, consider the observed treatment group average represents a combination of cases experiencing zero effects and threshold level effects, with proportions π and $(1-\pi)$, respectively. This gives us: $\bar{X}_T = (0+\bar{X}_C)\cdot\pi + (Thr+\bar{X}_C)\cdot(1-\pi)$, from which we can solve for π: $\pi = 1 - (\bar{X}_T - \bar{X}_C)/Threshold$. In this scenario, π represents the proportion of null effect cases that need to be replaced with threshold level effect cases to sustain the inference. A larger π indicates more evidence is needed to sustain the inference, which means the estimated effect is further away from the threshold for making an inference.

1.6.3 Case replacement for within-site spillover effects.

Now we want to relax the assumption of no spillover effects: the experiment condition of one unit can affect the outcome of another unit. That is, we want to conceptualize the potential bias due to spillover effects within each site so that we can characterize how sensitive the inference of each randomized control trial is to potential bias caused by spillover effects.

It is important to note that, in order to generate bias, the spillover effects should not be introduced by the treatment condition, otherwise it becomes a *mediator* that should be counted as part of the total treatment effect. We argue that spillover effects should satisfy at least two conditions to introduce bias: (1) inhere in direct interactions or indirect exposures among individuals; (2) perform in a way that is independent of the treatment assignment.

As before, consider we observe sample averages of treatment and control groups, denoted as $\bar{X}_T$ and $\bar{X}_C$, respectively, and we have $\bar{X}_T - \bar{X}_C > Threshold$ for a significantly positive intervention effect. But we wonder if the observed difference $\bar{X}_T - \bar{X}_C$ might be biased by either positive spillover effects in the treatment group or negative spillover effects in the control group, or even both.

Specifically, assume in the treatment group (with sample size n_T) we have a proportion (represented by π_T) of cases teaching the other treatment cases $n_T \cdot (1 - \pi_T)$. For each case in the teaching group, they benefit from teaching the other cases. Define the positive effect of teaching one case as TE (teaching effect), then teaching other $n_T \cdot (1 - \pi_T)$ cases gives each of them the benefit of $n_T \cdot (1 - \pi_T) \cdot TE$. Similarly, each case of the learning group can experience positive learning effect of $n_T \cdot \pi_T \cdot LE$ through studying from the other $n_T \cdot \pi_T$ cases. Define the isolated average outcome in the treatment group, without the spillover effects, as $\bar{X}_{T_T}$, then we should be able to get the following formula, representing that the observed treatment average is the sum of the isolated treatment average and positive spillover effects:

$$\begin{aligned} \bar{X}_T &= \bar{X}_{T_T} + \pi_T \cdot n_T \cdot (1 - \pi_T) \cdot TE + (1 - \pi_T) \cdot n_T \cdot \pi_T \cdot LE \\ &= \bar{X}_{T_T} + n_T \cdot \pi_T \cdot (1 - \pi_T) \cdot (TE + LE) \end{aligned}$$

Similarly, assume in the control group (with sample size n_C) we have a proportion (represented by π_C) of cases distracting the other control cases $n_C \cdot (1 - \pi_C)$. Then each first group case experiences a level of $n_C \cdot (1 - \pi_C) \cdot SID$ self-initiated distraction effect and each second group case experiences a level of $n_C \cdot \pi_C \cdot PID$ peer-initiated distraction effect. Denoting the isolated average outcome in control group as $\bar{X}_{C_T}$, we should be able to get:

$$\begin{aligned} \bar{X}_C &= \bar{X}_{C_T} - \pi_C \cdot n_C \cdot (1 - \pi_C) \cdot SID - (1 - \pi_C) \cdot n_C \cdot \pi_C \cdot PID \\ &= \bar{X}_{C_T} - n_C \cdot \pi_C \cdot (1 - \pi_C) \cdot (PID + SID) \end{aligned}$$

Now we can write out the formulas for isolated treatment and control group average, without any spillover effects, as follows:

$$\bar{X}_{T_T} = \bar{X}_T - n_T \cdot \pi_T \cdot (1 - \pi_T) \cdot (LE + TE)$$

$$\bar{X}_{C_T} = \bar{X}_C + n_C \cdot \pi_C \cdot (1 - \pi_C) \cdot (PID + SID)$$

By subtracting the second equation from the first, we get the difference between $\bar{X}_{T_T}$ and $\bar{X}_{C_T}$ as the treatment effect. Setting this effect equivalent to the threshold so that the amount of bias can invalidate the inference, we can get:

$$\bar{X}_T - \bar{X}_C - Threshold = n_T \cdot \pi_T \cdot (1 - \pi_T) \cdot (LE + TE) + n_C \cdot \pi_C \cdot (1 - \pi_C) \cdot (PID + SID)$$

This illustrates a general situation when the bias comes from both positive spillover effects in the treatment group and negative spillover effects in the control group. We can also consider two special situations: 1) all the bias comes from positive spillover effects in the treatment group (i.e., $PID + SID = 0$) and we can get $\bar{X}_T - \bar{X}_C - Threshold = n_T \cdot \pi_T \cdot (1 - \pi_T) \cdot PE$, where $PE(\textit{positive effect}) = LE + TE$; 2) all the bias comes from negative spillover effects in the control group (i.e., $LE + TE = 0$) and we can get $\bar{X}_T - \bar{X}_C - Threshold = n_C \cdot \pi_C \cdot (1 - \pi_C) \cdot NE$, where $NE(\textit{negative effect}) = PID + SID$.

In all three situations, we see the difference between the estimated treatment effect $\bar{X}_T - \bar{X}_C$ and the threshold is written as either a linear function of spillover effect (i.e., PE or NE), given π and n, or a quadratic function of π given the spillover effect (i.e., PE or NE or both) and n. To simplify the discussion, take the scenario where all bias comes from positive effects from the treatment group as an example. Figure 1.1 shows bias introduced from spillover effects as a quadratic function of π_T at different levels of spillover effect (i.e., PE). The axis of symmetry is $\pi_T = 0.5$. Consider the situations when $0 < \pi_T < 0.5$. If $\pi_T > 0.5$, we can always find another π_T from $(0, 0.5)$ that generates the same $(\bar{X}_T - \bar{X}_C - Threshold)$ by symmetry. Then given a fixed amount of PE, we can see that a larger π_T indicates a larger difference between the observed estimated effect and the threshold. That is, we need to have stronger spillover effects (more individuals teaching others in the treatment group) to invalidate the inference, indicating a more robust inference. Similar conclusions hold for sites whose estimated effect is so large that the distance to the threshold $(\bar{X}_T - \bar{X}_C - Threshold)$ is larger than the maximum value of this function, which is $0.25 n_T PE$.

Another way to interpret this is to assume 50% of cases teach the other 50% of cases in the treatment group (i.e., $\pi_T = 0.5$). Then we can solve for $PE = TE + LE$. That is, PE tells us how large the positive spillover effects must be to invalidate the inference of a positive intervention

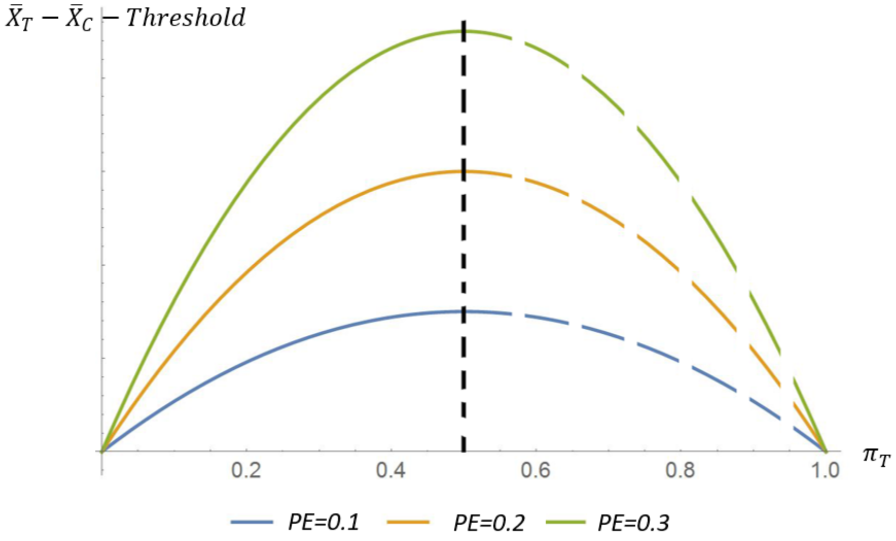


Figure 1.1: Bias introduced by spillover effects.

effect? The larger the resulting PE is, the more robust the inference is to the threat of spillover effect.

Additionally, by assuming $\pi_T = \pi_C = \pi$ and $PE = NE = R$, we can further simplify the formula for the general scenario where bias is introduced by both positive spillover effects in the treatment group and negative spillover effects in the control group: $\bar{X}_T - \bar{X}_C - Threshold = (n_T + n_C) \cdot \pi \cdot (1 - \pi) \cdot R$. Then we can apply the same approach above, using either π or R to describe the robustness of inference to potential spillover effects as bias. This way we can reduce the number of sensitivity parameters while considering spillover effects in both treatment conditions. Underlying the discussions for spillover effects are also counterfactual interpretations. By removing the spillover effects among participants, we are in fact trying to come up with the estimated effect when the SUTVA assumption is satisfied and individuals are independent from each other. Alternatively, we can consider replacing individuals with other individuals who experience the same treatment effect but no spillover effect.

The discussion above has made several assumptions for spillover effects. To illustrate, we use Table 1.1 to present the relations between senders and receivers of spillover effects indicated by the discussion above, when we wonder whether the estimated effect is biased by positive spillover

Table 1.1: Quantifying robustness of inference to potential spillover effects under restricted assumptions.

Individuals within treatment group		Teaching		Learning	
		A	B	C	D
Teaching	A	NA	B → A = 0	C → A = TE	D → A = TE
	B	A → B = 0	NA	C → B = TE	D → B = TE
Learning	C	A → C = LE	B → C = LE	NA	D → C = 0
	D	A → D = LE	B → D = LE	C → D = 0	NA

effects within the treatment group. Assume four individuals are in the treatment group: A, B, C, and D. 50% of them teach the other 50% ($\pi_T = 0.5$). Specifically, consider A and B teach C and D. Then Table 1.1 displays the specific spillover effect experienced by each pair of individuals based on the discussion above. For example, in the first row, A experiences TE (teaching effect) by teaching C and D but A does not experience any spillover effects from B. Similarly, C experiences LE by learning from A and B but does not experience any spillover effects from D. As such, several assumptions are implied here. First, we assume spillover effects are present for pairs of individuals within one treatment condition group (i.e., either treatment or control) but across teaching and learning groups (or self-initiated distracting and peer-initiated distracting groups). Second, each type of spillover effect (either LE, TE, SID or PID) is constant for different individuals. Inspired by the Linear-in-Means model in the peer effect literature, these assumptions can help us simplify the sensitivity approach technique and we can easily apply this for any unknown types of spillover effects.

But what if we have specific spillover effects that violate these assumptions? Then a weighted matrix of relations between senders and receivers of spillover effects can provide us with a more powerful and flexible tool to realize any possibilities. For simplicity, still assume we have four individuals: A, B, C and D. But now A and B are from the treatment group and C and D are from the control group. Then we can use a four by four table, as presented in Table 1.2, to specify spillover effects between each pair of individuals, whether they are from the same experimental condition or not. For example, the first row lists the spillover effects experienced by individual A from individuals B, C and D. Importantly, all the off-diagonal cells can have different values (the diagonal elements are meaningless because they represent one experience spillover effect from himself/herself). Assume we have weighted friendship data for all individuals in the sample, then we can use this table to ask how strong the unit spillover effect through the friendship network needs to be to invalidate the inference.

Table 1.2: A general approach for quantifying robustness of inference to potential spillover effects.

		Treatment		Control	
	Inidividuals	A	B	C	D
Treatment	A	NA	B → A	C → A	D → A
	B	A → B	NA	C → B	D → B
Control	C	A → C	B → C	NA	D → C
	D	A → D	B → D	C → D	NA

Now consider a simple application of this weighted matrix for a scenario where we have a positive but insignificant estimated treatment effect (i.e.,$Threshold > \bar{X}_T - \bar{X}_C > 0$). We wonder if the observed difference $\bar{X}_T - \bar{X}_C$ might be downward biased by positive spillover effects from the treatment to the control group. That is, individuals in the control group experience positive effects by interacting with individuals in the treatment group. Table 1.3 demonstrates how Table 1.2 can be applied to study this scenario, where only the bottom left panel shows spillover effects because that panel represents spillover effects from all individuals from the treatment group to all individuals in the control group.

Table 1.3: Quantifying robustness of inference to potential spillover effects from treatment to control group.

		Treatment		Control	
	Inidividuals	A	B	C	D
Treatment	A	NA	0	0	0
	B	0	NA	0	0
Control	C	A → C	B → C	NA	0
	D	A → D	B → D	0	NA

Now define the positive learning effect experienced by each individual in the control group through interacting with one individual in the treatment group as LE, each control individual gets an amount of $n_T \cdot LE$ positive spillover effects. Denote the isolated control mean as $\bar{X}_{C_T}$, then we can write: $\bar{X}_{C_T} = \bar{X}_C - n_T \cdot LE$. Now by setting the difference between treatment ($\bar{X}_T$) and isolated control mean ($\bar{X}_{C_T}$) equivalent to the threshold, we obtain: $Threshold - (\bar{X}_T - \bar{X}_C) = n_T \cdot LE$. From this we can calculate how large LE must be to sustain a positive treatment effect inference. A larger difference between the threshold and the estimated treatment effect indicates larger positive spillover effects (i.e., LE) must be present to sustain an inference.

1.6.4 Case replacement for within-site heterogeneous treatment effects.

As illustrated by Holland (1986), the average causal effect is an average and thus it "enjoys all of the advantages and disadvantages of averages". The constant treatment effect assumption says that all the units in the population of interest experience the same treatment effect caused by the treatment. This assumption will then allow the average treatment effect to be used to draw causal inference at the unit level.

One may argue that we only need to draw an inference at an average level rather than at a unit level. But in educational MSTs, each site may only have very small sample size. The number of participants in either treatment or control group can be even smaller. For example, small classes in Project STAR may include fewer than 20 students and most schools only had one to two small classes. With such small sample sizes, an outlier can have a considerable effect on the group average.

To quantify how sensitive the inference about the intervention effect in each site is to the heterogeneous or the outlier effect, we propose a successive extreme replacement thought experiment. Because the outlier effect can affect either the treatment or the control group or both, this selective replacement approach can be applied to both groups. To illustrate, consider a scenario where we have $\bar{X}_T - \bar{X}_c > Threshold$ for a significantly positive intervention effect. The goal of this discussion is to characterize how robust this inference for a specific site is to outlier effects. Under this scenario, we may consider outliers from the higher end in the treatment group or lower end in the control group. Specifically, we start our replacement from the individual who has the most extreme outcome in the group, which means the highest in the treatment and the lowest in the control group under this scenario. For replacement cases, we may consider the overall grand mean outcome across all participants within this site, favoring the null hypothesis that there is no treatment effect. Alternatively, we can replace extreme cases in each group with their own group mean, which means replacing the highest in the treatment with the treatment mean and replacing the lowest in the control with the control mean. Figure 1.2 shows both approaches, where the blue dotted line represents the control distribution and red solid line represents the treatment distribution. If replacing the most extreme case is not enough to cross the threshold, we continue by replacing the second extreme case. We continue this process until the difference between the treatment and control group reduces to the threshold and we record how many cases need to be replaced to invalidate the inference. The more we need to replace, the more robust the inference is.

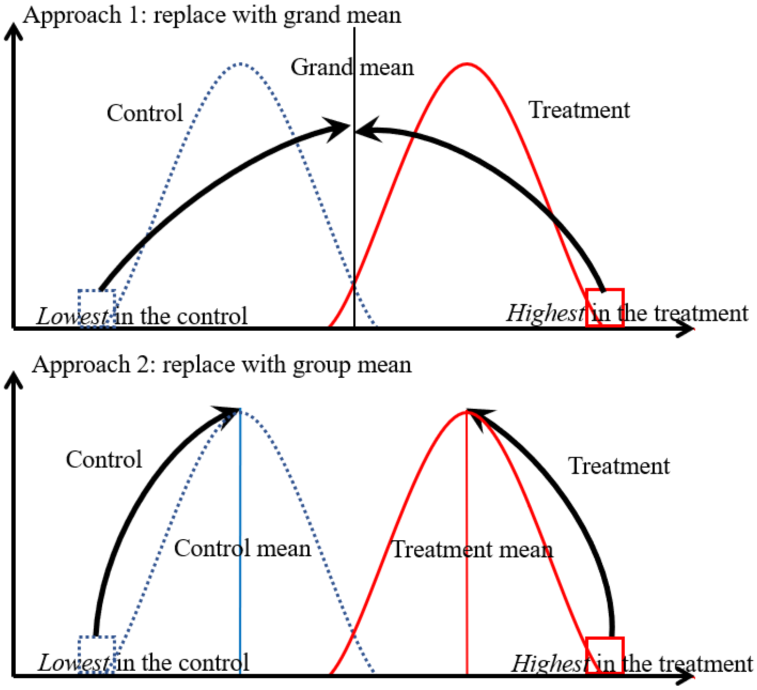


Figure 1.2: Successive extreme replacement for within-site heterogeneous treatment effects.

1.6.5 Heterogeneous treatment effects across sites.

All previous discussions focus on randomized control trials within each site. But one important advantage of the MST is that it allows researchers to study how consistent the intervention effect is across different sites. To characterize this cross-site variation, we apply the case replacement approach for each single site and consider four groups of sites. Figure 1.3 presents how this may work. The first group (Group 1) includes all sites that show a significantly positive intervention effect. Applying our case replacement approach, we can quantify the strength of evidence for each site by considering what percent of observed cases need to be replaced with no effect cases to invalidate the inference of a positive intervention effect. As such, we get a proportion π_j for each site j, which allows us to generate a distribution of robustness represented by π_j. The more sites with larger π_j (towards 1) the stronger the evidence for a positive intervention effect across sites. The second group (Group 2) includes all sites that show a positive but not significant effect within site, under which the case replacement approach allows us to have π_j indicating how much more evidence we need to sustain a positive effect inference. Then the more sites with a smaller π_j (towards 0), the more evidence is given towards a positive intervention effect. The third group

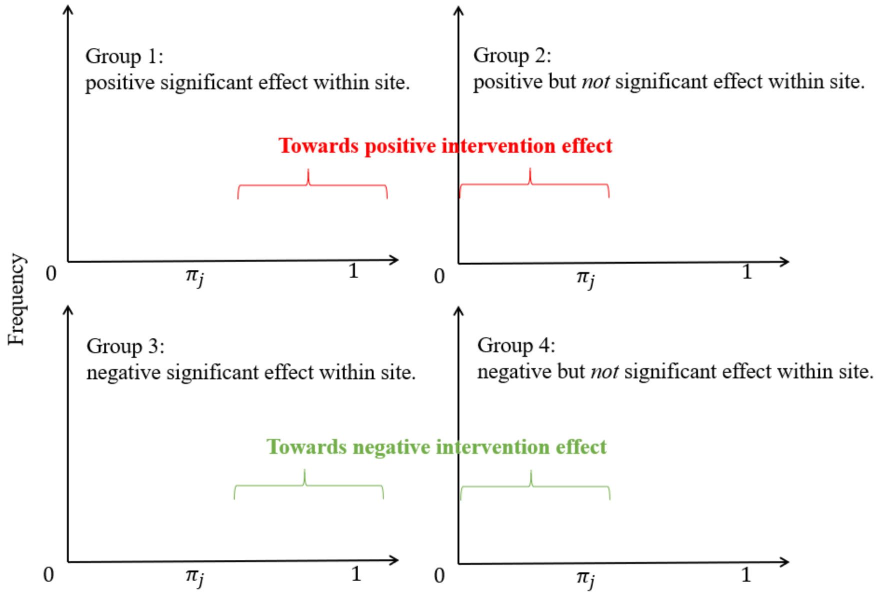

Frequency
Group 1:
positive significant effect within site.
Group 2:
positive but *not* significant effect within site.
Towards positive intervention effect
0
π_j
1
0
π_j
1
Group 3:
negative significant effect within site.
Group 4:
negative but *not* significant effect within site.
Towards negative intervention effect
0
π_j
1
0
π_j
1

Figure 1.3: Cross-site variation of the intervention effect.

(Group 3) includes all sites with significantly negative within-site treatment effect. More sites showing a large π_j (towards 1) indicate stronger evidence towards a negative effect. The final group (Group 4) includes all sites with negative but not significant effects. The more sites showing small π_j (towards 0) indicate a trend towards a negative intervention effect.

More importantly, from how all sites are distributed across all four groups, we can see 1) where most sites are; 2) whether sites are distributed towards inferences of different directions of intervention effect; and 3) variation in the robustness of inference. For example, if there are many sites that are located in the center of both group 1, 2 and group 3, 4, as shown by both the green and red brackets, then we know that the variation of cross-site intervention effect is substantial as we have pretty robust evidence towards inferences of both positive and negative intervention effects.

1.7 Illustrative Example of the Study of Class Size Effect in Project STAR

In this section, we will use Project STAR as an example to demonstrate how the discussion above can be applied to quantify the strength of evidence in multisite randomized trials to potential bias. For simplicity, we will focus on the math achievement of students in Grade K who were assigned to either small classes or regular classes. We excluded those students in regular classes with a full-time aide so that we have two experimental conditions to be consistent with the previous discussion. In general, we consider the small class to be the treatment group and the regular class to be the control group. In total, there are 3,794 students from 79 schools included in the following analysis. We standardized the math achievement scores for each school to make the interpretation easier.

1.7.1 Case replacement when SUTVA holds.

We start with the general case replacement approach when the SUTVA assumption holds. Specifically, we look at two schools as examples, both of which have a significantly positive treatment effect but having very different levels of robustness of inference. Table 1.4 shows the summary statistics of math achievement by class type for each school. Because students and teachers are

Table 1.4: Summary Statistics of Math Achievements by Class Type in School A and B.

Math achievement		*N*	*M*	*SD*	*min*	*max*
School A	Small class	13	1.321	0.794	0.183	3.171
	Regular class	34	-0.426	0.735	-1.761	1.037
School B	Small class	14	0.498	0.941	-1.286	2.002
	Regular class	23	-0.154	0.916	-2.153	1.314

Notes. *N*, *M* and *SD* represent sample size, mean, and standard deviation, respectively. *min* and *max* represent minimum and maximum values, respectively.

randomly assigned to each class type, we applied an independent samples t-test to estimate the small class effect[1]: the estimated effect for school A is 1.738 with a p value < 0.001 and the estimated effect for school B is 0.652 with a p value of 0.045. If we apply the commonly used threshold of 0.05, we may conclude that both schools show a significantly positive class size effect.

But do they have the equal strength of evidence or the same level of robustness of inference? To consider this, we apply our case replacement framework, which tells us: around 71.63% of the cases in school A must be replaced with null effect cases to invalidate the inference while only 2.26% of the cases in school B must be replaced with null effect cases to invalidate the inference. This shows the inference of an effect in school A is much more robust than that in school B.

1.7.2 Case replacement for within-site heterogeneous treatment effect.

As shown above, the inference of a positive small class effect in school B is not robust. Now we use the selective replacement approach to see whether this inference is very sensitive to outlier effects. Following Figure 1.4 shows a distribution of math achievement by class type in school B. We observe that there are a few students showing very low math achievement in the control group (regular class). Specifically, the lowest score is -2.152. Once we replace this student (the blue part) with a hypothetical average student with the grand mean of 0.093 (the pink part), the average score in regular class increases to -0.056 and the corresponding difference from the average in small class reduces from 0.652 to 0.554, which is lower than the threshold of 0.637. This means only by

[1] This is essentially the same approach as Konstantopoulos (2011) used. The difference is that we excluded the full class with aide and accordingly, we did not use Bonferroni correction.

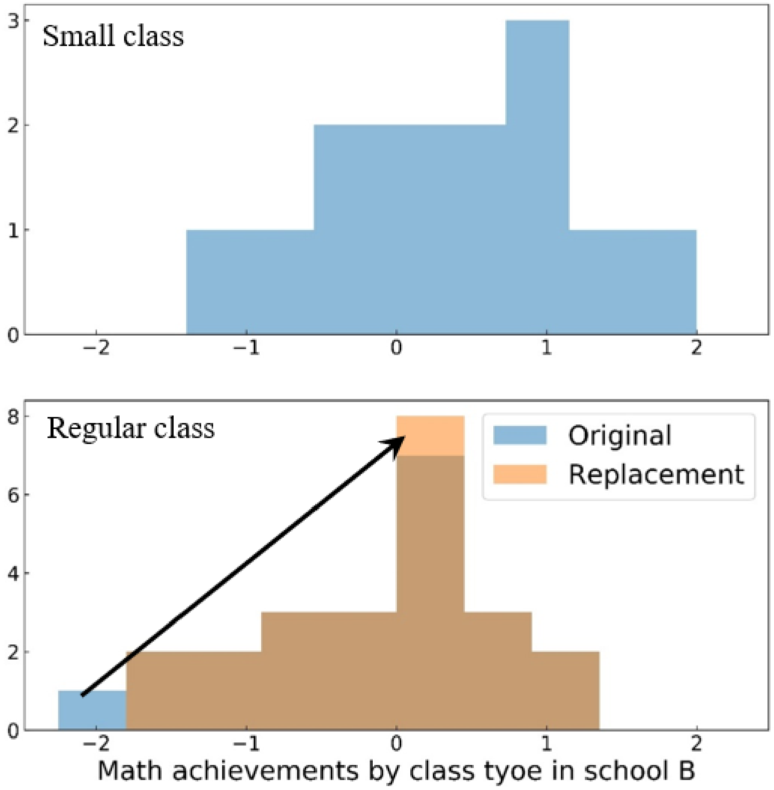


Figure 1.4: Math achievement by class type in school B.

replacing one student who has the lowest score in the regular class, the inference of a positive small class effect would be invalidated, indicating a very weak inference to potential outlier effects.

1.7.3 Case replacement for spillover effects.

Now we consider how sensitive the inferences in school A and B are to potential spillover effects. Assume 50% of students assigned to small classes teach other 50% of students in small classes. If all the bias comes from these positive learning and teaching effects, each unit of the spillover effect needs to be larger than 0.38 (more than 22% of the estimated treatment effect) in school A to invalidate the inference while in school B the positive spillover effect among students assigned in small classes only needs to be larger than 0.004 (less than 1% of the estimated treatment effect) to invalidate the inference. Alternatively, assume 50% of the students assigned to regular classes distract other 50% of students. If all the bias comes from these negative distraction effects among students who were assigned to regular classes for school A, the distraction effect needs to exceed 0.146 (more than 8% of the estimated treatment effect) to invalidate the inference while in school B, the distraction spillover effect only needs to exceed 0.0026 (less than 0.4% of the estimated treatment effect) to invalidate the inference. Again, this comparison between school A and B shows the large difference in terms of robustness of inference to potential bias.

We can also consider positive spillover effects from small class students to regular class students, leading to underestimation of the small class size effect. To illustrate this, we look at one school where 17 students were in the small class with an average math achievement of 0.264, and 20 students were in the regular class with an average math achievement of -0.423. The difference is 0.687, which is just below the threshold for statistical significance ($\alpha = 0.5$) of 0.707. In order to sustain an inference of a positive small class size effect, we can apply Table 1.3 as a tool to quantify how large the positive spillover effect from small class size to regular class must be. After calculation[2] , each student in the regular class must experience an amount of 0.0012 spillover effect from each student in small class, which is only about 0.17% of the estimated treatment effect, to alter the inference regarding a positive small class size effect. This quantifies the robustness of no effect in the school.

1.7.4 Case replacement for cross-site heterogeneity.

Figure 1.5 applies the case replacement approach to present cross-site variation in small class size effects on math achievement. First, many schools do not show any evidence of either positive or negative effects, indicated by the concentration of the distribution on the very right of Group 2 and Group 4. Second, many schools are in Group 1 or 2, indicating the estimated small class effects are positive. Additionally, 7 schools in Group 1 have a robust inference for positive small class size effects: more than 40% of observed cases need to be replaced by unobserved zero effect cases

[2] To sustain the inference, the average in regular class must be lower than $0.264 - 0.707 = -0.443$. Then the difference between this and the observed average in regular class is $-0.423 - (-0.443) = 0.02$. Each student in regular must learn at least 0.02 from all students in small class, then each student must learn $0.02/17 \approx 0.0012$, which is only about $0.0012/0.687 \approx 0.17\%$ of the estimated treatment effect.

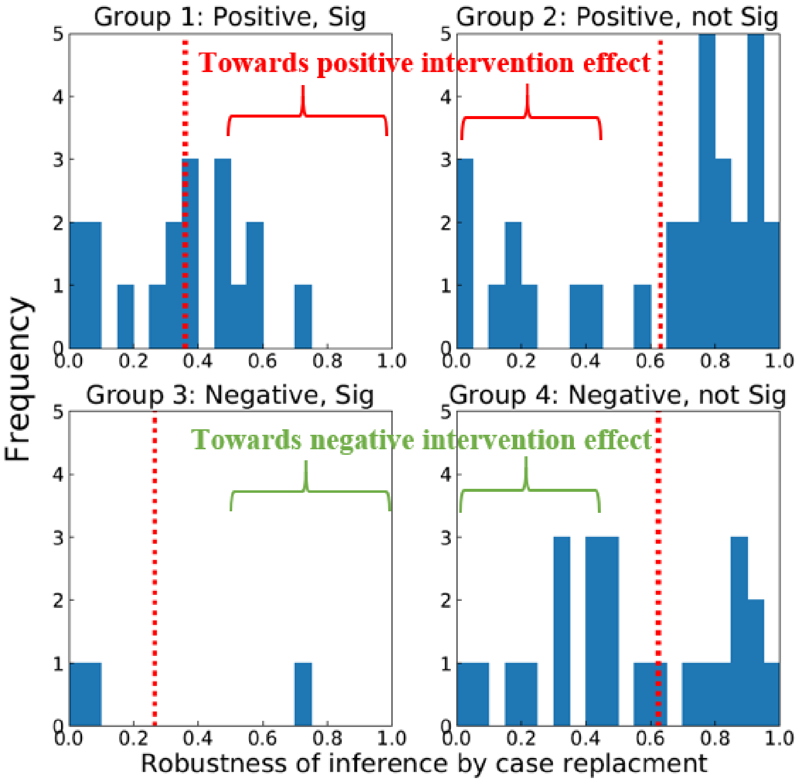


Figure 1.5: Cross-site variation in the robustness of inference of a small class size effect on math achievement (Grade K).

to invalidate the inference. Meanwhile, a few schools in Group 2 are close to the threshold: there are 3 schools in which fewer than 10% of zero effect cases need to be replaced with threshold level cases to sustain a positive intervention effect inference. However, as shown by the two frequency distributions for Group 3 and 4 in the second row, there are a few schools that show negative estimated effect, among which 3 are significantly negative. One school in Group 3 has such strong evidence to support a negative effect of small class size that one would need to replace more than 70% of the cases with zero-effect case to invalidate the inference. Additionally, in Group 4, a few schools are not too far away from the threshold of negative class size effects. Therefore, we may conclude that schools do differ from each other in terms of the inference of an effect of small class size on students' math achievement in kindergarten.

The red dotted lines in each group in Figure 1.5 represents the overall robustness for each group when school sizes are considered to weigh each school. That is, within each group, we summed up the number of students needed to be replaced from all schools and divided this by the total number of students. Figure 1.6 further aggregates all information into one figure: the distribution illustrates

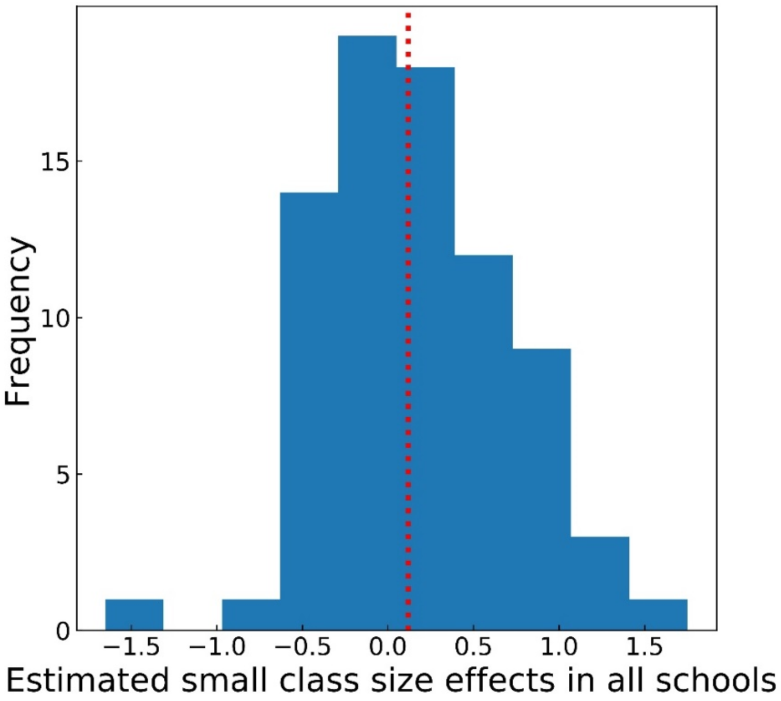


Figure 1.6: Summarizing small class size effect across schools in one Figure.

the estimated treatment effect across all schools, and the dotted line indicates the threshold for a positive treatment effect, for which we used the standard error as Konstantopoulos (2011) calculated for the average small class effect across schools. Comparing Figure 1.6 with Figure 1.5, we argue that Figure 1.5 provides much more detailed information about how inconsistent the small class effect is across schools. Although one estimate that summarizes all schools is appealing, missing the heterogeneity across schools can provide misleading information for policymakers. Moreover, summarizing all schools with one estimated effect ignores the fact that each school will make its own inference about the effectiveness of small classes. This applies to scale-up because schools have local control over policy. Thus, schools must ask if the intervention worked in their context.

To further show how our approach can provide richer information regarding cross-site heterogeneity in MSTs, we carried out a similar analysis for Grade 1 students regarding whether small class size influence their math achievement. Figure 1.7 presents the result. Comparing Figure 1.5

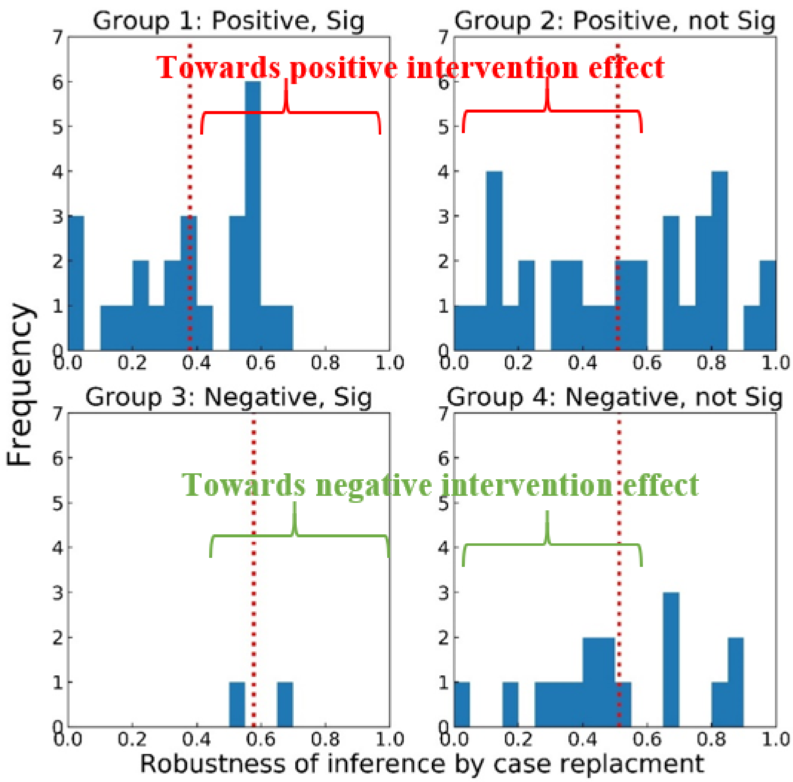


Figure 1.7: Cross-site variation in small class size effect on math achievement (Grade 1).

with Figure 1.7, we can tell the inferences regarding small class size effects on math achievement in Grade 1 and Grade K share both similarities and differences[3]. First, compared to Grade K, there are more schools in Grade 1 showing robust findings for positive small class size effects; in Group 1 of Grade 1, there are 11 schools for which more than 50% observed cases need to be replaced by unobserved zero effect cases to invalidate the inference. Meanwhile, both Grade 1 and Grade K have schools with strong evidence for negative effect as well. In Grade 1, two schools in Group 3 show strong evidence for negative small class size effects. In the strongest school more than 70% of the cases must be replaced with zero-effect case to invalidate the inference. Therefore, we may conclude that Grade 1 has even stronger evidence for positive small class size effects but both Grade K and Grade 1 show cross-site heterogeneity with evidence for both positive and negative intervention effects.

[3] The sample for Grade 1 includes 76 schools, compared to 79 schools for Grade K.

1.8 Discussion

This chapter extends the case replacement approach (Frank et al., 2013) to quantify strength of evidence in multisite randomized control trials, accounting for spillover and heterogeneity. Throughout, Project STAR has been applied as an example to demonstrate how this non-parametric approach can better inform debate regarding relevant policy choice. Drawing on Rubin's causal model (RCM) (Rubin, 1974), concerns about bias are expressed in terms of unobserved, counterfactual data. Specifically, the robustness of a causal inference is interpreted as the percentage of a sample that must be replaced with counterfactual no-effect cases to alter the inference. Most importantly, by carefully considering how to select cases from the observed sample to be replaced, what cases are regarded as replacement cases, and whether the non-replaced cases in the sample experience any changes during the replacement, this case replacement framework can be extended to attend to violations of the Stable Unit Treatment Value Assumption (SUTVA) and presence of heterogeneous treatment effects.

One limitation of this chapter is that the case replacement approach might be better motivated when an overall inference across sites is based on the accumulation of inferences within each site. In that case the inference within each site must be compared with the site-specific threshold as I have done here. Additionally, the discussion of spillover effects can be more thorough and better tailored for the context of multisite randomized control trials, considering that the randomization can help eliminate some type of spillover effects but not others. For example, assuming randomization is implemented with fidelity, we should not expect any presence of within treatment group (or control group) spillover effects due to non-random assignment of distractors or contributors. However, spillover effects between different treatment groups cannot be excluded by randomization, especially in the context of educational interventions, where the subjects are always students who interact with each other in a social context. While these issues are present in any study, they are especially prominent when we compare inferences across sites, as in the multisite trial.

CHAPTER 2

QUANTIFYING STRENGTH OF EVIDENCE FOR INFERENCES IN VALUE ADDED MEASURES: CASE REPLACEMENT, SPILLOVER, AND HETEROGENEITY

2.1 Introduction

The introduction establishes the goal of understanding the robustness of inferences in terms of how the sampling mechanism could be altered to replace cases in the data. In the previous chapter, we derived the case replacement approach to quantify strength of evidence for inferences in multisite randomized control trials, accounting for spillover effects and heterogeneous treatment effects. In this chapter, we will continue to demonstrate how this case replacement framework can be applied to quantify uncertainty in value-added measures (VAMs) that have been used to evaluate teacher effectiveness. The VAM context is highly relevant for policy and personnel decisions because the evaluation of teacher effectiveness is related to high-stake decisions. By comparing students' expected test scores to their actual ones, the "deflections" are inferred to be the "added value" from the teacher (Raudenbush & Bryk, 2002). Proponents of value-added models cite research that shows teachers' considerable and long-lasting influences on student achievement (e.g., Chetty, Friedman, & Rockoff, 2011; Hill, Kapitula, & Umland, 2011; Rivkin, Hanushek, & Kain, 2005). They argue that there is important variation in teachers' effectiveness that can be better identified by VAM (e.g., Aaronson, Barrow, & Sander, 2007; Hanushek & Rivkin, 2010). By selecting or deselecting teachers based on value-added we can improve teacher quality and increase student achievement and long-term outcomes (e.g., Chetty et al., 2011; Gordon, Kane, & Staiger, 2006; Winters & Cowen, 2013a, 2013b). For example, under the evaluation system of IMPACT in the District of Columbia Public Schools (DCPS), teachers were dismissed with rare exceptions once evaluated as "ineffective" and the dismissal threats are claimed to have helped improve teacher performance and student achievement (Adnot, Dee, Katz, & Wyckoff, 2017; Dee & Wyckoff, 2015).

Various concerns have been raised about the validity and reliability of value added measures as a basis to inform teacher evaluation (e.g., Guarino, Reckase, & Wooldridge, 2015; Harris, 2009; Raudenbush, 2015). These issues may include test unreliability, missing data and model misspecifications. In order to get a more reliable value added measure, researchers recommend model specification with two years of prior tests (Goldhaber & Hansen, 2010; Kane & Staiger, 2012; Rothstein, 2009). This can reduce the population of teachers who can be measured because it is not uncommon for teachers to change grades or students to change schools within three-year duration. Furthermore, unreliability in test scores can appear as measurement errors or differences among different achievement measures. Previous research has shown bias in value-added caused by measurement errors alone (Lockwood, Louis, & McCaffrey, 2002) as well as large variation in the estimated effects of applying different achievement measures (Lockwood et al., 2007). Therefore, those high-stake decisions for a specific teacher (e.g., hiring, retention, and professional development) require all stakeholders to be able to understand and conceptualize the uncertainty and quality of the measures to avoid unfairness and loss of investments and resources.

2.2 Spillover and Heterogeneity in Value-added Measures (VAMs)

Like MSTs, VAMs also demonstrate typical education scenarios where SUTVA is rarely satisfied in practice. In the VAM context for teacher evaluation, SUTVA suggests there are no peer effects that can affect a student's achievement, which is rarely satisfied in real life. Previous researchers have controlled peer effects through model specification (e.g., Carrell, Fullerton, & West, 2009; Hoxby & Weingarth, 2005; Levin, 2002; Van Ewijk & Sleegers, 2010). For example, the most well-known peer effects are based on classmates' achievement levels. Other peer effects can be generated by racial, ethnic or economic forces. Unfortunately, identifying all possible sources of peer effects is challenging as it would have to include non-cognitive attributes as well as interaction styles. As a result, we will never know whether we have controlled for all potential significant peer effect mechanisms in a model. For example, how can we control the peer effects from friendships if we do not have adequate information on friendship to define closer knit peer effects more than simply class membership?

In the VAM context, heterogeneous treatment effects are present if one teacher is good at teaching certain students but not others, which can cause debate in teacher evaluation. Additionally, the VAM context exemplifies the scenario where violations of one single treatment (as a fundamental component of SUTVA) may lead to heterogeneity of treatment effects, as discussed in Chapter 1. For example, teachers may modify their way of teaching to better suit the special requirements of each student, which may lead to heterogeneous teacher effects experienced by different students in one classroom. Even with the same teaching approach, different students can still benefit differently due to variations in their aptitudes, motivations, prior knowledge or any other individual or contextual factors.

2.3 Strength of Evidence in VAMs

In VAM, the threshold for being an effective teacher is generally arbitrary since the estimation of standard errors in VAM is controversial; the standard errors can be quite sensitive to how one conceptualizes the level of analysis and what formula one chooses. Regardless of the specific definition/calculation, a threshold in VAM represents the point at which the evidence from VAMs would make one indifferent to the final teacher evaluation result, such as effective or ineffective. As argued in Frank et al. (2013), the comparison between the threshold and the VAM then represents the strength of evidence that supports the evaluation result that directly links to high-stake personnel decision-making. Thus, all stakeholders, including teachers, administrators and parents, should be able to understand this comparison so that the personnel decision-making can be made after comprehensive consideration and evaluation of the strength of evidence against potential costs. For example, consider a debate between an administrator and a teacher whose VAM is below the threshold of being effective. The administrator may use this below-threshold VAM to evaluate this teacher as ineffective, but the teacher may argue that her low VAM is primarily because she has been assigned to lower-end students. Essentially, this debate is about how far the teacher's VAM is below the threshold, and whether this difference generates strong evidence for this teacher's lack of effectiveness and potential dismissal, considering all potential sources of bias in estimating VAM.

As we will demonstrate in this chapter, the case replacement approach proposed in Frank et al. (2013) provides an intuitive alternative to quantify the comparison between the threshold and the estimated VAM to inform discussions among all stakeholders. Additionally, it does not rely on the controversial calculation of standard errors.

Like Chapter 1, this chapter has two goals. First, we want to demonstrate how the case replacement approach can be applied in VAMs. Second, we aim to extend Frank et al. (2013) work by discussing how we can quantify the uncertainty to violations of SUTVA and presence of heterogeneity in VAMs. In the following section, we will first review the case replacement framework proposed by Frank et al. (2013).

2.4 Case Replacement as a Counterfactual Thought Experiment

In Frank et al. (2013), the authors showed an approach to quantify how much bias there must be in an estimate to invalidate an inference. Then the bias is interpreted in terms of sample replacement to inform more intuitive interpretation. In other words, to show how robust an inference is, we ask a question based on a thought experiment which is counterfactual: what percentage of the sample should be replaced with counterfactual (unobserved) no-effect cases to invalidate an inference made from the data? Or if the concern is about external validity, we consider what percentage of the sample should be replaced with no-effect cases from an unsampled population. The larger the percentage is, the more robust the conclusion/inference is, the less likely that the finding is only due to chance or bias.

This case replacement idea can be applied in various ways to characterize the strength of evidence in VAMs. But the general idea is always about replacing some students taught by one teacher with counterfactuals of other students, if they were taught by that teacher, as a thought experiment. The replacement process can become very flexible depending on: (1) how we select students from the teacher's class to be replaced (2) what students are regarded as replacement students (3) whether the non-replaced, remaining students in the class experience any changes in achievement during the replacement. In the following analysis, we will discuss different ways to consider the hypothetical replacement and how each approach helps inform the comparison between the threshold and estimated VAM under different contexts.

2.5 Case Replacement for Quantifying Uncertainty in VAMs

2.5.1 Sources of bias in VAMs.

Various concerns have been raised about the validity and reliability of value added measures (VAMs) as a basis to inform high stake decisions (e.g., hiring, retention, and professional development) for a specific teacher (e.g., Guarino et al., 2015; Harris, 2009; Raudenbush, 2015). We will begin our review of these concerns with the conditional random assignment assumption. Almost all the potential inconsistencies of value added, such as those caused by test unreliability, missing data or model specification, can be represented in terms of the violations of conditional random assignment assumption. Our approach to quantify the uncertainty of value-added also draws on this framework of the student-teacher assignment mechanism.

The conditional random assignment assumption is that students are randomly assigned to every teacher conditional on the other variables (Rothstein, 2009, 2010). However, research has shown that there is a nontrivial amount of sorting based on students' prior test scores as well as a nontrivial amount of non-random assignment of teachers to classrooms. For example, recent research has shown that teachers who are nominated as help-providers to other teachers and with leadership positions are assigned better students (Kim et al., 2018). These nonrandom assignments may cause substantial bias in value added estimates if not captured by the controls in the model specification (Paufler & Amrein-Beardsley, 2014; Rothstein, 2010). In some cases, the estimates based on value added may even have the opposite sign of the true teacher effect (Dieterle, Guarino, Reckase, & Wooldridge, 2015). Hiring or dismissing a teacher based on this flipped ranking can be unfair for teachers and cause unwanted competition that can lead to test-driven teaching.

2.5.2 Case replacement when SUTVA holds.

To better motivate the case replacement approach, we start our discussion with a hypothetical example. Consider three teachers' VAMs as presented in following Figure 2.1. Both Ashley and Jessica are below the threshold of 0.15. If the threshold represents a serious lack of effectiveness, the administrator may decide to dismiss both of them. However, we can see that Ashley is much closer to the threshold than Jessica. This indicates that an evaluation of Ashley as ineffective is much less robust than that for Jessica. It might be some bias in VAM estimation causes teacher Ashley to be below the threshold. As a result, the personnel decision should be considered more seriously, or other measures should be referred to. Or the administrator may direct Ashley to professional development if the school resources can only support one teacher for this opportunity. Similarly, an administrator may want to provide professional development to teacher Emily who is above the threshold, but just barely so.

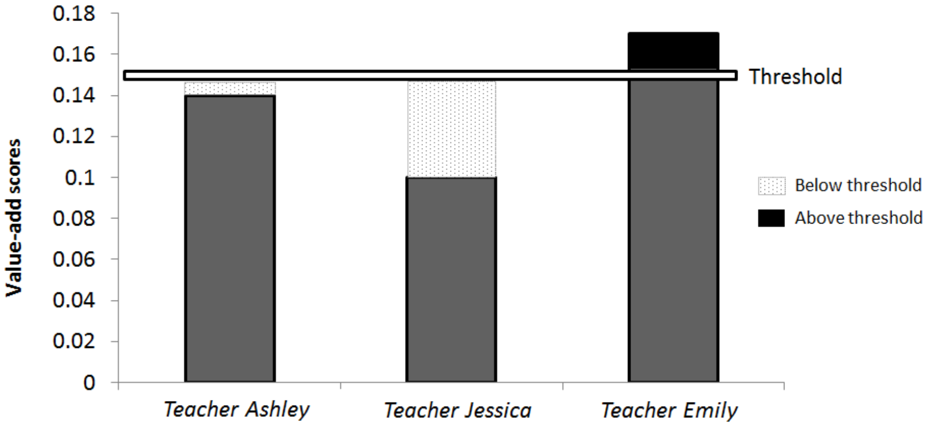


Figure 2.1: Teacher effects estimated by VAM (hypothetical example).

In the hypothetical example, how can we better quantify the difference between Ashley's VAM and the threshold as strength of evidence for her evaluation? The case replacement approach leads us to ask how many students in her class need to be replaced with average students to get her to the threshold. Ashley has a VAM of 0.14, which could be the average of her students' gain scores. The threshold is 0.15. Assume she has 20 students, whose gain scores have a distribution represented as black and grey parts shown in 2.2. Hypothetically, to improve Ashley's VAM from 0.14 to 0.15, we can replace two students (the grey parts) with two average students whose gain scores are 0.16 (the white parts with black outline). This counterfactual thought experiment tells us that via replacing the two worst students with grade-average students, Ashely could achieve the threshold of being effective. We can also say that 2 out of 20, that is about 10% of Ashley's students need to be replaced with grade average students to alter the evaluation.

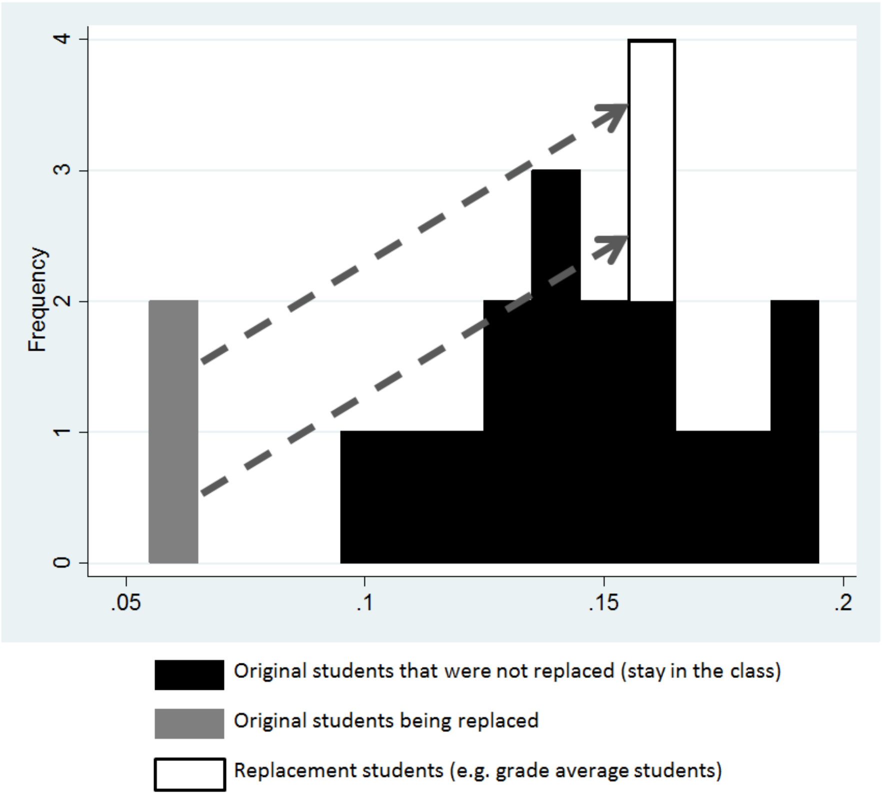


Figure 2.2: Example replacement of students to invalidate Ashley's evaluation based on VAM.

We now formalize the intuition in 2.2. For the following discussion, assume that all grade nine math teachers in one middle school were evaluated based on their students' achievement scores. Suppose we have a general value-added model as follows:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + T_{it}\gamma + X_{it}\beta + \mu_{it},$$ [1]

[1] There are various value added models. This particular (simplified) model is only used as an

where A_{it} is student i's test score at time t (post-test score); τ_t is the intercept; λ is the coefficient (scaler) for the pre-test score $A_{i,t-1}$; $A_{i,t-1}$ is student i's test score at time $t-1$ (pre-test score); T_{it} is a row vector of teacher indicators;[2] γ is a column vector of teacher fixed-effects;[3] X_{it} is a row vector that include covariates to control student heterogeneity such as student family backgrounds; β is a column vector that include the coefficients for the covariates X_{it}; μ_{it} is an unobserved error term.

After estimating those parameters, we obtain a "purified" gain score s_{il} for each student i in teacher l's class at time t after removing the effects from those observed characteristics ($A_{i,t-1}$ and X_{it}) included in the value-added model. This is shown in the following equation:

$$s_{il} = A_{it} - (\hat{\tau}_t + \hat{\lambda} A_{i,t-1} + X_{it}\hat{\beta})$$

This gain score s_{il} can also be understood as a "deflection" score which is the difference between a student's expected score (based on those covariates $A_{i,t-1}$ and X_{it} that are outside teacher l's control) and the actual score. We assume that this deflection is caused by teacher l who teaches student i.[4] To clarify, all the gain scores in following discussions refer to this s_{il}. We can decompose s_{il} by using ANOVA parameterization:$s_{il} = \mu + \alpha_l + e_{il} = VAM_l + e_{il}$, where μ is the grand mean gain score of all students in this grade. For simplicity, we assume that each teacher teaches one

example to illustrate that: all the following discussions are based on the "purified" or adjusted "gain scores". In other words, the gain score here is after adjusting for student characteristics that are included in the value-added model. Theoretically, the change in students' test scores can be decomposed to teacher effectiveness and student heterogeneity. By removing the latter part, we can estimate the teacher effectiveness.

[2] If there are 10 teachers in this grade, then T_{it} is a 1×10 row vector for each student (observation). Correspondingly, γ is a 10×1 column vector, with each element representing one teacher's fixed effect.

[3] Here "fixed-effect" means that we are NOT viewing all teachers as a population and then getting an estimate based on a random sample drawn from this population of teachers. Instead, we are interested in learning individual teacher's effect on student achievement. Therefore, we just use dummy variables to indicate each teacher. There are also other estimation methods in value added literature. For example, when we apply Project STAR later to illustrate our approach, we will demonstrate two approaches: the EB residual approach and the Dynamic OLS approach.

[4] Here we contextualize our discussion by using student-level data to evaluate teachers within a school as an example. Therefore, the school-level effect is not included.

class and each class has the same number of students n, then $\mu = \overline{VAM}$, which is the average VAM of all teachers in this grade; α_l is how far teacher l's value added score (VAM_l) departures from the $\overline{VAM}$; VAM_l[5] is the VAM for teacher l; e_{il} is how far the score of student i in teacher l's classroom departures from the classroom mean.

To simplify the discussion for now, assume that we are evaluating teachers for one grade within one school. One way to set the threshold is to use a certain percentile such as the 5^{th} percentile in all teachers' VAM distribution.

For teacher l, we can only observe her effect on the students in her class. For the other students taught by other teachers, we cannot know their scores if they were taught by teacher l because this is counterfactual. Because of this, teacher l may argue that her value-added VAM_l is below the threshold (Thr) because of the students she is assigned. She may argue she has the average teacher effect and she will be able to achieve the threshold if she is assigned more grade average students (this could well be the argument of a beginning teacher – see Kim et al. (2018)). However, the evaluator, such as the principal, may argue that the teacher's low VAM_l reflects teacher l's lack of effectiveness. While the dispute is about the point estimate of the VAM, the debate about the teacher's evaluation is informed by understanding the uncertainty of the VAM.

To formalize the discourse above for the uncertainty of value-added, the case replacement approach leads us to consider how many grade average students need to replace to alter teacher l's evaluation. Another argument for replacing with grade-average students is that if we randomly choose one student from the grade then a grade-average student will be the expectation for a student being selected.

In order to carry out the replacement analysis, we need to know the grade average student's gain score. As before, this gain score is achieved after adjusting for all those covariates included in the value-added model. Two possible ways are presented as follows to get an estimate for this grade-average student's gain score g_t.

In the first approach, we can just use μ as an estimate for g_t. This approach is convenient and

[5] If the pre-test score (test score for time 1) is set before the teacher encounters the student, then we can think about this VAM as a function of the post-test score (test score for time 2).

the resulting g_t will be the same for all teachers in the replacement thought experiment. From the teacher's argument illustrated before, she has the average teacher effect in this grade and this g_t might be a good estimate for an average teacher's effect on a grade average student. The disadvantage is that this average gain score is under the observed teacher-student assignment condition and we are assuming that this grade average student will keep the grade average score if taught by this teacher.

Another possible way to estimate this grade average student's gain score g_t is still conditioning on covariates in the current model and the observed teacher-student assignment but is more conservative. Specifically, rather than look for an estimate for an average teacher's effect on a grade average student, we try to estimate this particular teacher l's effect on a grade average student. The "average" here refers to having the grade average pre-test scores and other controlled characteristics. This means looking for a student j in teacher l's class so that the value of $\left|\left(A_{j,t-1} + X_{jt}\right) - \ (\bar{A}_{t-1} + \bar{X}_t)\right|$ is minimized (where $\bar{A}_{t-1}$ and $\bar{X}_t$ are the grade average covariates). Then we use this student j's gain score s_{jl} as g_t. The second approach here seems to provide a "closer guess" for a grade average student's gain score if taught by teacher l. However, since we only use one specific student's observed gain score for the replacement, the reliability will be a more serious issue than the first approach.

Once we get g_t, we consider randomly selecting students from teacher l's class to be replaced with grade average students, then the formula is shown as follows.

$$Thr = (1 - \pi) \cdot VAM_l + \pi \cdot g_t = (g_t - VAM_l) \cdot \pi + VAM_l$$

From this we can get: $\pi = \frac{Thr - VAM_l}{g_t - VAM_l} = 1 - \frac{g_t - Thr}{g_t - VAM_l}$, where π is the percentage of students need to be replaced, Thr is the threshold of value-added above which the teacher will be evaluated as effective. In this chapter, we assume that the threshold (Thr) is below the average value added score ($\overline{VAM}$) and all the VAM_l we are interested in is below the threshold (Thr). That is, we have $VAM_l < Thr < \overline{VAM}$.

Suppose g_t is bigger than Thr and VAM_l. Also the g_t is the same for all teachers (the first approach discussed previously). Then with higher VAM_l, the π gets smaller. This makes sense intuitively because teachers who are closer to the threshold need to replace fewer students. However, we can see that the relationship between VAM_l and π is not linear.

If we treat π as a function of VAM_l (and assume g_t as a known constant for now), we can apply delta method to get a standard error for π as follows.

$$\frac{dVAM_l}{d\pi} = (Thr - g_t) \cdot (g_t - VAM_l)^{-2}$$
$$(\frac{dVAM_l}{d\pi})^2 = (Thr - g_t)^2 \cdot (g_t - VAM_l)^{-4}$$

Then we get:

$$AVar[\sqrt{n}(\hat{\pi} - \pi)] = (Thr - g_t)^2 \cdot (g_t - VAM_l)^{-4} \cdot AVar[\sqrt{n}(\widehat{VAM}_l - VAM_l)]$$

From this formula, we note that as the VAM moves further away from the grade average gain score (that is, the $(g_t - VAM_l)$ gets larger), the asymptotic variance of $\hat{\pi}$ approaches 0 because of the term $(g_t - VAM_l)^{-4}$. This indicates that for a teacher with a relatively low VAM, we can get a quite precise estimate for π. In other words, the more certain we are as the VAM gets further away from the threshold.[6] In that case, a relatively large $\hat{\pi}$ can represent a quite robust evaluation for a teacher's ineffectiveness relative to a threshold.

In order to account for randomness in estimating $\hat{g}_t$, we consider a bootstrap approach. To generate one bootstrap sample, we resample students with replacement within each teacher. Then for each such sample, we can calculate $\widehat{VAM}_l$, $\hat{g}_t$ and $\hat{\pi}$. Repeating this step allows us to get a distribution of $\hat{\pi}$ where a confidence interval can be obtained that accommodates both randomness and potential bias. One advantage of this approach is it accounts for the uncertainty of $\widehat{VAM}_l$ via bootstrap as well, without going into debates about what formula we should use to calculate the standard error. If this confidence interval is close to 0, then the evaluation for this teacher as being ineffective is sensitive to either potential bias or sampling variability or both.

Similar arguments can be applied for scenarios where the estimated VAM is above the threshold: $Thr < \widehat{VAM}_l$. Now the goal is to quantify how the data must change to get this teacher below

[6] Note the discussion here assumes the uncertainty of VAM (i.e., $AVar[\sqrt{n}(\widehat{VAM}_l - VAM_l)]$) keeps unchanged.

threshold because we wonder whether the teacher's above-threshold VAM might be due to upward bias and if so, the teacher may need some professional development. The case replacement approach here is similar to that in Chapter 1 when we wanted to quantify how the data must change to sustain an inference. Specifically, consider the estimated VAM represents a combination of students with grade average gain scores and threshold level gain scores, with proportions π and $(1 - \pi)$, respectively. This gives us: $VAM_l = g_t \cdot \pi + Thr \cdot (1 - \pi)$, from which we can solve for π: $\pi = \frac{VAM_l - Thr}{g_t - Thr}$. In this scenario, π represents the proportion of students with grade average gain scores that must be replaced with students who have threshold level gain scores to bring this teacher below the threshold. A larger π indicates more data must be changed to change the teacher evaluation to ineffective, which means the estimated VAM is higher and further away from the threshold.

2.5.3 Selective case replacement for heterogeneous effects.

As we discussed for MSTs, we can apply the case replacement approach to quantify how robust a teacher evaluation inference is to outlier students in his/her class. Note in the VAM context, this constant effect assumption does not necessarily relate to student grouping based on prior test results. Prior test scores may reflect students' ability but the constant effect assumption is about teachers' effects on students. Students' ability may or may not relate to their improvement affected by the teachers. The treatment effect in this context is more a problem of whether this teacher's teaching works for one student (matching problem). Therefore, even in the most homogeneous case where students are grouped based on their pre-test scores, we still need to think about violation of the constant effect assumption.

Another important note here is the heterogeneous teacher effects can be caused by violations of SUTVA, or more specifically, the single treatment level assumption in SUTVA. This happens when a teacher modifies their teaching for different students. But as discussed earlier, the presence of heterogeneity of teacher effects on student achievement does not rely on violations of this single treatment level assumption. Whether and how one teacher's teaching benefits student learning can be affected by various factors.

This constant effect assumption is also highly relevant in the VAM context: if we want to rank all teachers, we should consider the heterogeneity of the students in their classes and ideally the teacher whose teaching works out for more students should be more favored. Consider two teachers who have the equivalent value-added. In teacher l's class, there is only one student who gets an extremely low gain score and it is this score that makes the teacher's value-added below the threshold. However, teacher m has several students who get quite low gain scores. In addition to the estimated value-added, we can also use this information of mismatching as another measure for teacher's effectiveness. Even if we are only interested in the average level, we may still be concerned about the effects of outliers.

To quantify how sensitive the VAM is to this heterogeneous or the outlier effect, we propose three selective replacement approaches as follows, assuming the teacher has a below-threshold VAM.

(1) Successive extreme replacement: this process is data-dependent and there is no closed formula. This is very similar to what we discussed in the context of MSTs: we start our replacement from the student who has the lowest gain score in teacher l's class. If the teacher's value-added is still lower than the threshold, then we replace the student with the second lowest gain score. We continue this process until teacher l's value-added achieves the threshold and we record how many students need to be replaced with g_t.

(2) Purposeful sampling process: the lower the student i's gain score is, the higher the probability for this student gets selected to be replaced. This is shown in the following formula:

$$For\ all\ s_{il} < VAM_l, \Pr\left(s_{il}\ is\ selected\ to\ be\ replaced\right) = \frac{VAM_l - s_{il}}{\sum\left(VAM_l - s_{il}\right)}$$

Then the formula for replacement is shown as follows:

$$Thr = \sum_{s_{il}<VAM_l}\left[\left(g_t - s_{il}\right)\cdot\frac{VAM_l - s_{il}}{\sum\left(VAM_l - s_{il}\right)}\right]\cdot\pi + VAM_l$$

From this we obtain:

$$\pi = \frac{Thr - VAM_l}{\sum_{s_{il} < VAM_l} \left[(g_t - s_{il}) \cdot \frac{VAM_l - s_{il}}{\sum (VAM_l - s_{il})} \right]}$$

For now we are only replacing students who are below the class average (for all $s_{il} < VAM_l$). But we can also consider including those students who are above the class average but below the threshold ($Thr > s_{il} > VAM_l$). In this case, the formula will be the following one.

$$\pi = \frac{Thr - VAM_l}{\sum_{s_{il} < Thr} \left[(g_t - s_{il}) \cdot \frac{Thr - s_{il}}{\sum (Thr - s_{il})} \right]}$$

(3) Replace the teacher's median student(s) with grade average students. This approach is proposed considering that median is less sensitive to extreme values than mean. Instead of replacing the mean student gain score in the teacher's class, we select the median student to think about the replacement (Med_l). The formula is represented as follows.

$$\pi = \frac{Thr - VAM_l}{g_t - Med_l}$$

For any of these approaches, the magnitude of π gives us an intuitive understanding about how far teacher l is from the threshold if we assume the value-added is a valid and reliable evaluation. It also quantifies how much bias there needs to be to invalidate this evaluation. A small π indicates a lack of robustness or a small departure from the threshold. Similarly, we may get a confidence interval for π by applying the delta method or bootstrap.

Additionally, the three selective replacement schemes can help provide a supplemental measure for teacher evaluation. For instance, when two teachers have the same value-added, we can use π from selective replacements as another measure for evaluation purpose. The teacher with a smaller π may be favored because her VAM is more likely to have been negatively affected by just a few outlier students.

2.5.4 Case replacement for peer effect (violation of SUTVA).

In this section, we will conceptualize the potential bias due to peer effects in value added models in terms of random or purposeful resampling of students in a counterfactual scenario. To start, we will discuss how peer effects generate bias in VAMs.

In the VAM context, we prefer to use "peer effects" rather than "spillover effects" because spillover can occur through teachers (such as one naughty student distracts the teacher's attention from the rest of the class). But we define peer effects as those direct effects among students. These peer effects are then just like other factors such as students' own background that we should control in value-added models.

To further define peer effects as a potential bias in value-added, we need first consider the baseline peer effects. The value-added model is essentially a normative comparison among teachers. Therefore, the baseline peer effects that are present in all teachers' classes should not be credited to one teacher. That is, to generate "bias" in value-added measures, peer effects should satisfy several conditions: (1) inhere in direct interactions or indirect exposures among students; (2) perform in a way that is independent of particular teachers; (3) is unique for a particular classroom that is not covered by the normative baseline peer effects. Specifically, the second requirement indicates that the peer effects do not depend on which teacher teaches the class, while the third requirement illustrates the existence of a biasing peer effect needs to be different from baseline peer effects that are experienced by all teachers. Assume we are comparing all the teachers within a school. Then the baseline peer effect depicts what happens in all teachers' classrooms in that school. But some peer effects are unique for the classroom. Consider teacher Ashley's classroom in which peer effects are not due to Ashley's instruction. These effects can bias our evaluation for Ashley.

It is also important to note that the level of baseline peer effects can be quite different in various school environments. Students involved in project-based learning can experience strong peer effects through frequent group discussions while students in schools with conventional teaching styles may only experience a minimal level of peer effect by observation. Therefore, we should consider these variations when discussing peer effects as potential bias in different contexts.

The non-random assignment of students to teachers can introduce bias to VAM via non-random assignment of contributors and distractors to teachers. For example, a teacher can be assigned with more students who are really distractors that may have negative effects on other students in the class. This negative spillover effect can introduce bias to the VAM. To quantify how sensitive one teacher's value-added is to potential bias due to such peer effect, we propose an approach that is very similar to what we discussed in Chapter 1 for spillover effects in the MST context. Consider a group of students $n\pi$ distracting the other students $n(1-\pi)$ in the class and they experience self-initiated (SID) and peer-initiated distraction effects (PID), respectively. The derivation is the same as the MST context. Following we present the result for the VAM context when we worry about negative peer effects as bias that may threat teacher's evaluation result. The notation is the same as before and VAM_{l_T} indicates the isolated/true VAM score for teacher l.

$$VAM_l = VAM_{l_T} - n\pi \cdot (1-\pi) \cdot SID - (1-\pi) \cdot n\pi \cdot PID$$

By setting VAM_{l_T} equivalent to the threshold of being an effective teacher, we can solve for π. As before, we only need to specify the negative effect $NE = SID + PID$ rather than specifying the self-initiated and peer-initiated distracting effects separately. Figure 1.1 in Chapter 1 still applies to show the relationship among the NE, π and the difference between teacher's VAM and the threshold (i.e., $Thr - VAM_l$).

$$Thr - VAM_l = -n \cdot NE \cdot (\pi - 0.5)^2 + 0.25 \cdot n \cdot NE$$

As mentioned before, underlying these discussions for potential peer effects are counterfactual interpretations. The hypothetical example presented in following Figure 2.3 illustrates this counterfactual idea in the VAM context, based on our previous example about teacher Ashley.

Recall Ashley has 20 students, where the first figure represents her 20 students' distribution of gain scores before replacement. Then in the thought-experiment, we replace the students indicated by the green part with the other students represented by the red part. Importantly, these two groups of students have comparable gain scores, but they have different peer effects on the remaining students: the green students will distract others but the red students only have baseline peer effects on others. Therefore, after the replacement, the remaining students in the class (represented by the black part) will not experience the peer-initiated distraction effects anymore. As such their scores get higher, causing the change in the teacher's VAM.

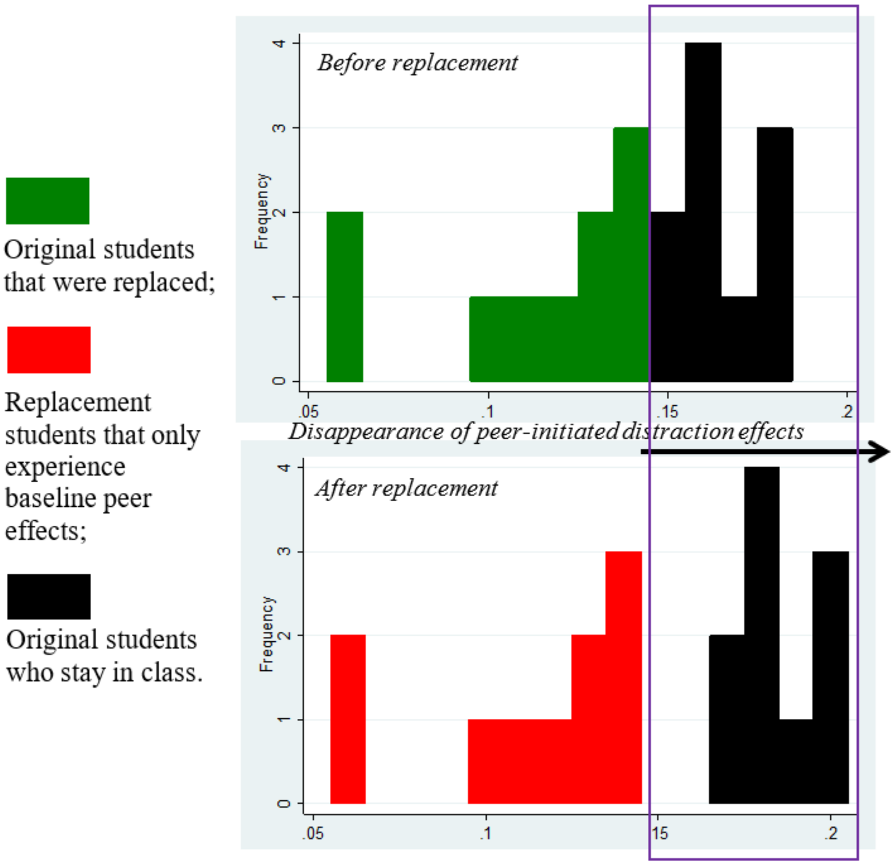


Figure 2.3: Case replacement for peer effects.

Importantly, the crucial trick in this thought experiment for peer effects is the remaining students change their gain scores during the replacement. When there is no peer effect, the change of teacher l's value-added is only from the difference between the new students' and original students' gain scores after the replacement. For example, in the hypothetical example in Figure 2.2, teacher Ashley needs $(0.15 - 0.14) \times 20$ $(the\ teacher\ has\ 20\ students) = 0.2$ total increase to achieve the threshold. This 0.2 comes from the difference between the two original students in the class (denoted in grey parts) and two replacement students (represented as the white parts), i.e., $(0.16 - 0.06) + (0.16 - 0.06) = 0.2$. Those original students who are not replaced do not change their scores in the replacement process. In contrast, in Figure 2.3, the gain in the VAM occurs because of the change experienced by the students remaining in the classroom.

Finally, 1.2 in Chapter 1 can again be applied as a more flexible framework for quantifying more specific peer effect mechanism if we know specific social interaction patterns within the teacher's classroom.

2.6 Illustrative Example of Evaluating Grade 1 Math Teachers in Project STAR

We will use Project STAR as an example to demonstrate how the discussion above can be applied to quantify uncertainty in VAMs. Different from the MST context, Project STAR is not designed for teacher evaluation purpose. The discussion here is mainly for demonstrating our case replacement approach. But as previous research illustrated (Nye, Konstantopoulos, & Hedges, 2004), the randomization of teacher assignments within schools and the broad range of schools from throughout a diverse state make Project STAR a great resource to study teacher effect variance on student achievement.[7] Specifically, because both students and teachers were randomly assigned to different classes, any systematic between-classroom variance in achievements should be due to either the treatment effects (class types) or teacher effectiveness. As such, in this paper, we first follow the approach applied in Nye et al. (2004) to study teacher effectiveness. Specifically, schools are included in our demonstration sample if there were more than three classrooms in the same grade so that within each school at least two classrooms were assigned to the same class type.[8] A three-level hierarchical linear models is applied, specified as follows.

Level 1 (student *i*)

$$Y_{ijk} = \beta_{0ij} + \beta_{1jk} Pretest_{ijk} + \beta_{2jk} Female_{ijk} + \beta_{3jk} FRL_{ijk} + \beta_{4jk} Minority_{ijk} + \varepsilon_{ijk}$$

[7] It is important to distinguish between: evaluate specific teachers versus evaluate teacher effect variance.

[8] Nye et al. (2004) has shown that the constrained sample is very similar to the complete sample on important characteristics.

Level 2 (teacher/classroom j)[9]

$$\beta_{0jk} = \pi_{00k} + \pi_{01k} Small_{jk} + \pi_{01k} Aide_{jk} + r_{0jk}$$

$$\beta_{1jk} = \pi_{10k}, \beta_{2jk} = \pi_{20k}, \beta_{3jk} = \pi_{30k}, \beta_{4jk} = \pi_{40k}$$

Level 3 (school k)

$$\pi_{00k} = \gamma_{000} + \eta_{00k}$$

$$\pi_{10k} = \gamma_{100}\ , \pi_{20k} = \gamma_{200}, \pi_{30k} = \gamma_{300}, \pi_{40k} = \gamma_{400}$$

At the student level, we control pretest, gender, whether the student is eligible for free and reduced lunch and whether the student is minority. At the classroom/teacher level, the treatment condition is controlled, and teacher random effect is applied. That is, r_{0jk} is the teacher effect for teacher j at school k. At the school level, the random component η_{00k} captures school k's effect on student's achievement score.[10] The Empirical Bayes estimates for the teacher random component r_{0jk} are then regarded as the estimated VAM for teacher j in school k.[11]

For simplicity, we focus on math achievements in Grade 1. As such, we get a sample with $3,209$ students taught by 268 teachers in 54 schools.[12] After fitting the three-level model illustrated above, we used the Empirical Bayes estimates for the 268 teachers as their VAM scores.

[9] For purpose of evaluating individual teachers, it might be better to exclude students in the regular class with an aide since in those classrooms, we cannot distinguish between teacher's effect from the aide's effect. Here we include these students in our sample to better align with earlier research and also this is only for demonstration of the case replacement approach.

[10] We decided to follow Nye et al. (2004) to treat teacher and school effect as random effects here for two considerations: (1) teacher fixed effect (i.e., teacher dummies) can be collinear with the treatment effect of different class types; (2) we acknowledge that school fixed effect can better control school effects on students but within each school there were only a few classrooms. Moreover, for most schools within each treatment assignment there were at most two classrooms, which can make the collinearity even worse. We also compared the coefficients of students' pretest, gender, SES and minority between school random effect and school fixed effect estimation. The results (both point estimate and standard error) are very similar to each other.

[11] Although we call teacher random effect and school random effect, this is different from random effect estimation in econometrics literature. In econometrics literature, random effect estimation in the VAM context mainly refers to student random effect when panel data is available. See Guarino et al. (2015) for more detailed information for different estimators for VAMs. What we use here is

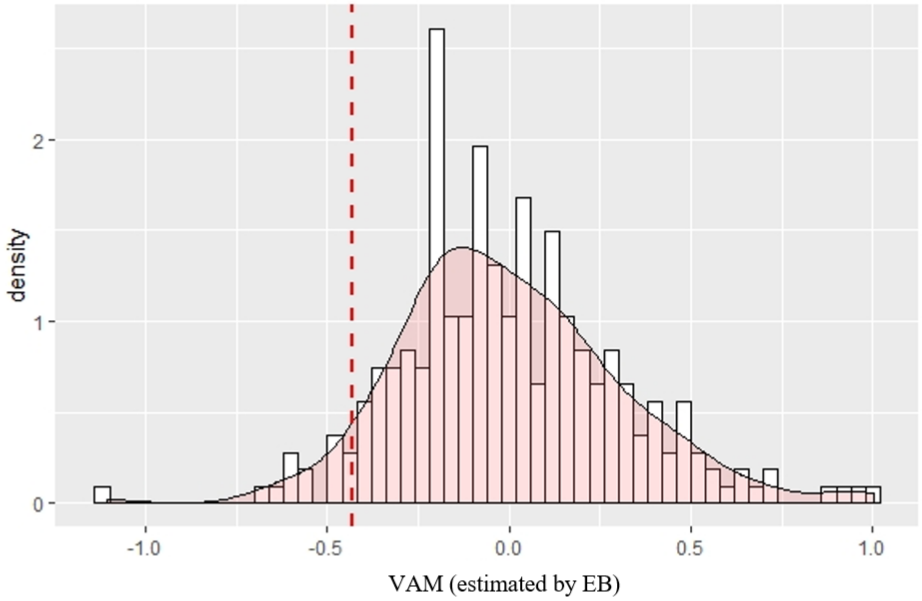


Figure 2.4: VAMs for 268 teachers.

Figure 2.4 shows a distribution of these VAM scores. Now consider the 5^{th} percentile of -0.43 as a threshold for teacher to be effective, indicated by the red dotted line in Figure 2.4. We choose this threshold because researchers have illustrated that student achievement in US can get to the level of Canadian by eliminating bottom $5-8$ percent of teachers (Hanushek, 2014). Additionally, in the controversial teacher evaluation system introduced in the District of Columbia Public Schools (IMPACT), teachers were dismissed with rare exceptions if they were evaluated as ineffective (bottom 3%) and researchers argued this dismissal threat increased performance of remaining teachers (Adnot et al., 2017; Dee & Wyckoff, 2015).

Based on this threshold of the 5^{th} percentile of -0.43, there are in total 14 teachers falling into this category of "being ineffective". Figure 2.5 presents the estimated VAMs for these teachers, where each blue bar represents VAM for teacher A through teacher N. Note the VAMs are negative, and a longer/lower bar indicates a worse estimate for teacher effectiveness. The red horizontal line

what they call "Empirical Bayes and Related Estimators".

[12] This sample is slightly different from that in Nye et al. (2004), but the results are very similar. This particular sample is selected based on several constraints: (1) school has at least four classrooms in grade 1 (2) students have both Grade K and Grade 1 math achievement score available, also not missing information about gender, race and whether the student was eligible for free and reduced lunch (3) we remove two teachers, for whom only one and three students' data are available.

represents the threshold of being evaluated as effective. Then the more the blue bar below the red threshold, the more robust teacher evaluation is.

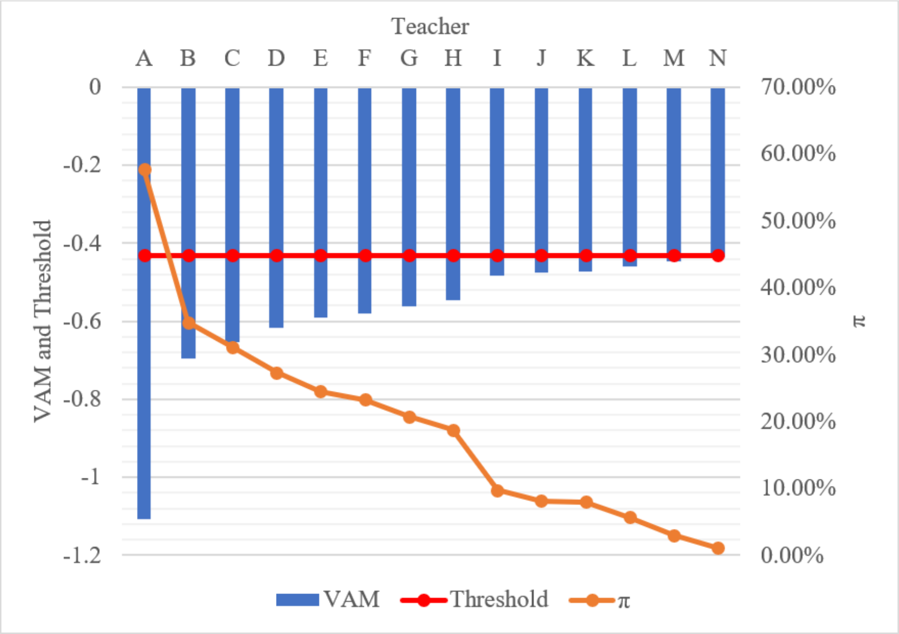


Figure 2.5: VAMs for 14 ineffective teachers and their robustness in terms of percent of students need to be replaced (π).

Now we apply our student replacement approach to quantify how robust the teacher evaluation is for these teachers. We first assume the SUTVA holds. To carry out the case replacement thought experiment, we first calculated the value of the replacement case (i.e., g_t in discussion above). In this context, the teachers are being evaluated against all the other teachers in this sample, and the argument made by an ineffective teacher is that they could achieve the threshold if they had been assigned with more average students in this sample. To approximate this counterfactual situation of being assigned with “an average student”, we calculated an average VAM weighted by each teacher’s number of students, to serve as an estimate for the teacher effect an average student can experience with an average teacher. As such we get an average of -0.0086.[13] Now we can apply the

[13] This is not too far away from the unweighted average of 0. We argue for this weighted version because in the counterfactual situation the teacher had been assigned with an average student and the weighted version captured this student assignment by focusing on the student level. But the results should be very similar no matter which one we use, as long as assignment of students to

student replacement approach to quantify how robust each teacher's evaluation is to any potential sources of bias (when SUTVA holds). The results are presented by the orange dots for each teacher. For example, for teacher A, more than 57% students need to be replaced with average students to get this teacher above the threshold. In comparison, teacher N only needs to replace a little more than 1% students in his/her class to get to the threshold. That is, the evaluation for teacher A is much more robust for the evaluation for teacher N. We ordered the 14 teachers in Figure 2.5 to have their VAMs getting closer and closer to the threshold from teacher A to N. Correspondingly, the percentage of students need to be replaced in the thought experiment also gets smaller. For teacher I through N, we only need to replace less than 10% of their students to get them to the threshold, indicating their evaluation is very sensitive to potential bias and accordingly, the related personnel decisions may need more considerations.

We acknowledge that the calculation of π (percentage of students need to be replaced) involves sampling uncertainty. Now we want to take into account the sampling variability for our index of inference robustness. To achieve this goal, we apply bootstrap approach to generate a distribution of π for each teacher. The bootstrap is based on within-teacher resampling with replacement. Following Figure 2.6 presents the result for teacher A and B (based on $1,000$ iterations). First all $1,000$ iterations generated a below-threshold VAM for teacher A. Additionally, as presented in Figure 2.6, the 95% confidence interval for this teacher's π is (42.09%, 60.93%). Thus, we may conclude that the evaluation for teacher A is pretty robust, even after we consider the sampling variability. In comparison, although teacher B also has a pretty large π of 34.82%, the sampling variability is much larger compared to that for teacher A. First there are 53 times out of $1,000$ iterations where B's VAM is actually higher than threshold. For the remaining 947 times, the 95% confidence interval is (3.88%, 46.12%), which is much closer to 0 compared to that for teacher A. This comparison shows teacher B's students are more diverse compared to teacher A, and thus teacher B is much more affected by sampling variability.

Now we want to relax the SUTVA and see how sensitive the teacher evaluation is to potential

teachers is not too unbalanced.

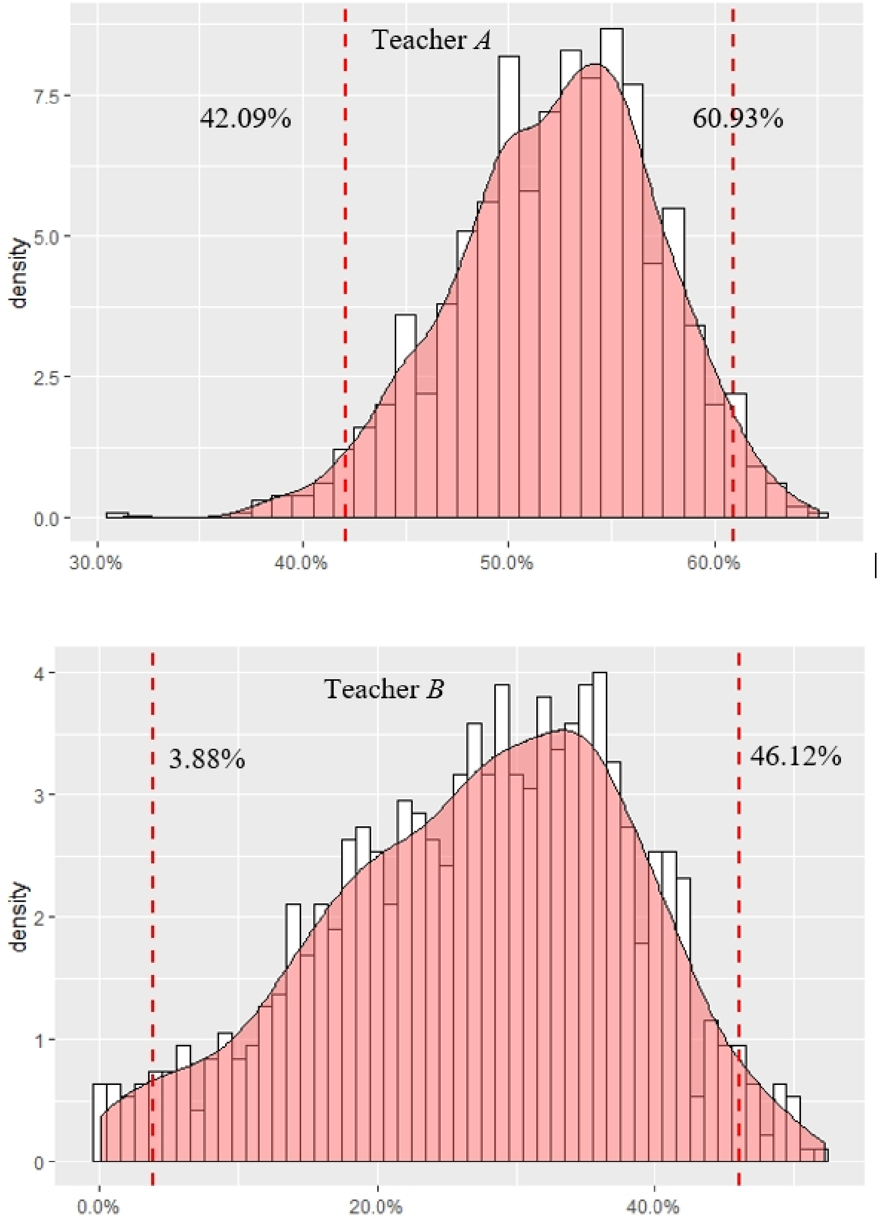


Figure 2.6: Sampling variability for percentage of students need to be replaced (π) for teacher A and B.

peer effects that can generate bias. Figure 2.7 presents the result by applying our approach for the 14 teachers below the threshold. As Figure 2.5, the blue bars indicate teachers' VAMs and the red horizontal line indicates the threshold. But now the green ones represent the smallest negative distraction effect to invalidate teacher evaluation if we assume 50% of the students are distracting the others in the teacher's class. The orange ones represent the distraction effect if we assume 10% of the students are distracting the others.

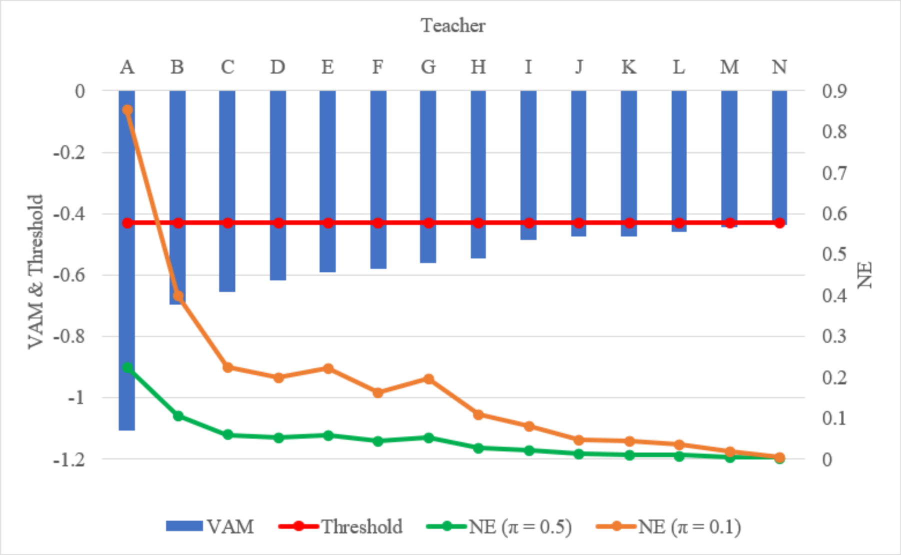


Figure 2.7: VAMs for 14 ineffective teachers and their robustness to potential peer effect as bias.

First, we observe assuming more students distracting others, smaller unit of distraction effects (NE) is needed to invalidate the inference (the orange dot is always above the green dot). As we discussed before, assuming 50% students distracting others gives us the smallest peer effect possible. Second, there is a difference between Figure 2.7 and Figure 2.5: in Figure 2.5, as teacher's VAM gets closer to the threshold, we always have smaller percentage of students need to be replaced to invalidate the teacher evaluation. But this is not always the case in Figure 2.7. For example, teacher G has a higher VAM than teacher F, but teacher G needs a stronger peer effect as bias to invalidate her evaluation than teacher F. This is because the calculation of peer effect takes into account the number of students in the class. This makes sense because peer effects essentially describe the interaction among students which should be related to the number of students. However, as we mentioned before, if one prefers more specific interaction styles, they can apply a more flexible approach presented in Table 1.2 in Chapter 1.

As mentioned before, we can always interpret Figure 2.7 in two ways: either interpreting the NE as the smallest necessary negative peer effect to invalidate the evaluation, or interpreting π as how many (more precisely, what percentage of) students need to be replaced with students having comparable scores but will only influence the other students with baseline peer effects. Take the teacher A for example. Assuming 50% of the students in the class were distracting others, then there must be at least 0.226 negative peer effects as bias to get the teacher to the threshold. What does 0.226 mean? It indicates that each of the 50% students (1^{st} group) are distracting each of the other 50% of students (2^{nd} group). And 0.226 is the sum of two effects: (1) one unit self-initiated distraction effect suffered by one student in the 1^{st} group by distracting one student in the 2^{nd} group; (2) one unit peer-initiated distraction effect experienced by each student in the 2^{nd} group due to being distracted by one student in the 1^{st} group. Because we standardized the outcome variables (the math achievements in Grade 1), the 0.226 means more than 2 standard deviations, which is a large effect. Alternatively, assuming the negative peer effect is 0.226, we can interpret the 50% as: 50% of the students need to be replaced with students who have comparable math achievements but only have baseline peer effects on their classmates. Therefore, each of the 50% remaining students can experience an increase of 0.226 (more than 2 standard deviations) in their math achievements once those distracting students get replaced. By applying the same bootstrap approach before, we can also get a 95% confidence interval for the negative effect, which is (0.147, 0.285) for teacher A. Additionally, consistent with earlier discussion, the sampling variability for teacher B's negative peer effect is also large. Among the 947 out of $1,000$ iterations where this teacher is below the threshold, the necessary negative peer effect has a 95% confidence interval of (0.010, 0.193).

The demonstration above uses the Empirical Bayes (EB) estimates for the VAMs. As is well known (e.g., Guarino et al., 2015), there are several different estimation methods for VAMs. In the following discussion, we apply a different estimation method, the Dynamic Ordinary Least

Squares (DOLS) approach, to estimate VAMs. We do this for two considerations. First, this DOLS approach better aligns with our derivation above and allows us to demonstrate the selective sampling for heterogeneity in an intuitive way. As pointed out by Guarino et al. (2015), the EB estimates are similar to shrinking teacher average residuals towards the overall mean, where the residuals are obtained after regressing posttest on pretest and other covariates, except teacher assignments. Since this shrinking process occurs at the teacher level and the magnitude of shrinkage can be different for each teacher, the VAM cannot be considered as a simple average of the students' adjusted gain scores any more, unless we are willing to conceptualize the shrinking procedure at the student level. But going deep into the shrinking procedure can complicate the thought experiment and make the sensitivity analysis technique less accessible to potential audience. Second, presenting two estimation approaches allows us to show how our student replacement approach works for different teachers and different estimation methods. We will show that the student replacement approach provides us with a general framework that can be easily applied to compare evaluation results generated from different estimation methods.

Specifically, we use the DOLS approach for a model specified as: $Y_i = \beta_0 + \beta_1 Pretest_i + \beta_2 Female_i + \beta_3 FRL_i + \beta_4 Minority_i + T_i\gamma + \varepsilon_i$, where T_{it} is a row vector of teacher indicators and γ is a column vector of teacher fixed-effects (i.e., VAMs). The other notation is the same as before, where Y_i is math achievement at Grade 1 and $Pretest_i$ is math achievement at Grade K. We restrict our sample to teachers who taught small classes so that we do not need to worry about collinearity between teacher effect and class type effect. As such, we get a smaller sample with $1,128$ students taught by 101 teachers. Figure 2.8 shows the distribution of these 101 teachers' VAMs (i.e., $\hat{\gamma}$ in the equation), where the red dotted line represents the 5^{th} percentile (-1.60) as a threshold for being an effective teacher. 5 teachers' VAMs are below this threshold.

Now we apply the student replacement approach to characterize the robustness of teacher evaluation based on VAMs for the 5 teachers who are below the threshold. Table 2.1 presents the results together with the robustness for the 14 teachers who are below the threshold in the EB approach. Specifically, the second column (# of students in total) reports the total number of

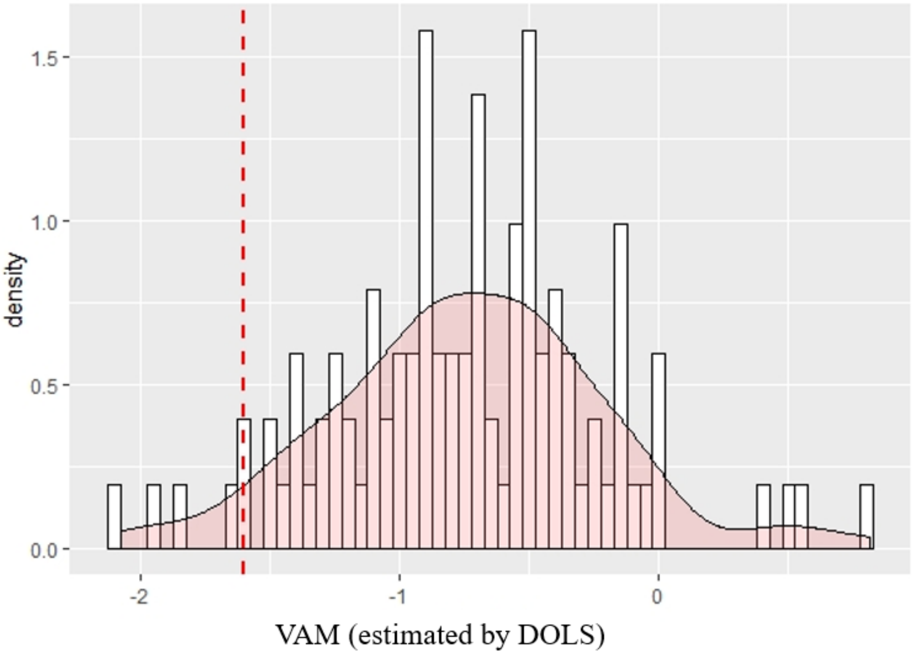


Figure 2.8: VAMs (estimated by DOLS) for 101 teachers who taught small classes.

students for the teacher. The third column (EB_π) reports the percentage of students that must be replaced to change the teacher evaluation in the EB approach. There are 5 ineffective teachers in the EB approach who did not teach small classes, thus the robustness of their evaluation in the DOLS approach is not applied (NA). The fourth column (DOLS_π) reports the percentage of students that must be replaced to change the teacher evaluation in the DOLS approach. In both the third and fourth column, if the percentage is within a parenthesis, then the teacher is above the threshold and the percentage indicates how much the data must change to get the teacher below the threshold. The last two columns (i.e., DOLS_het_π1 and DOLS_het_π2) present the results of two selective replacement approaches to examine the robustness of teacher evaluation to potential heterogeneous effects.

Table 2.1 allows us to see how our robustness index works for different teachers and different estimation methods. Note that the two approaches (i.e., EB and DOLS) are very different. The EB approach compares teachers from all treatment conditions (i.e., small class, regular class and regular class with aide). But the DOLS approach is only applied to teachers who taught small

Table 2.1: Case replacement approach for VAMs estimated by EB and DOLS.

Teacher	# of students in total	EB_π	DOLS_π	DOLS_het_π1 (# of students must be replaced)	DOLS_het_π2
A	12	57.74%	35.52%	33.33% (4)	26.93%
B	10	34.82%	1.56%	10.00% (1)	0.96%
C	15	31.14%	NA		
D	14	27.29%	NA		
E	11	24.52%	29.27%	18.18% (2)	14.29%
F	14	23.27%	(24.27%)		
G	10	20.75%	(0%)		
H	16	18.74%	(27.18%)		
I	10	9.73%	4.91%	10.00% (1)	1.99%
J	14	8.14%	NA		
K	15	7.95%	NA		
L	13	5.69%	(32.47%)		
M	12	3.02%	(10.94%)		
N	15	1.07%	NA		
P	8	(10.87%)	22.67%	25.00% (2)	16.70%

Note: The DOLS approach only includes teachers for teaching small classes.
NA indicates the teacher did not teach small class.

classes. All teachers in the DOLS approach are included in the EB approach, but not vice versa. As such, the two approaches generate different coefficients for the same predictors (i.e., students' gender, race and free and reduced lunch information), different VAMs, and different thresholds. But our student replacement approach allows us to compare the robustness of teacher evaluation results across these two approaches. Comparing the third and fourth columns (i.e., EB_π and DOLS_π), first we observe that four out of five ineffective teachers in the DOLS approach are also below the threshold in the EB approach: teacher *A*, *B*, *E* and *I*. In both approaches, we have strong evidence for teacher *A* to be ineffective. The only teacher that is below the threshold in EB but above the threshold in DOLS is teacher *P*. But our thought experiment tells us that in the EB approach, the evidence for teacher *P* being effective is not very strong since only 10.87% of average students must be replaced with threshold-level students to bring this teacher below the threshold. Similarly, our approach tells us that two estimation approaches give similar results in general for small class teachers who are ineffective in the EB approach but effective in the DOLS approach.

For example, teacher G is below the threshold in the EB approach and is just at the threshold in the DOLS approach (his/her VAM is exactly at the threshold of -1.60). Another example is: among the nine small class teachers below the EB approach threshold, the three teachers with the strongest evidence of being ineffective are also below the threshold in the DOLS approach (i.e., teacher A, B and E); it is the teachers with the weakest evidence in the EB approach that turn to be effective in the DOLS approach (i.e., teacher L and M). This indicates the VAMs generated by different approaches can be highly correlated. However, when it comes to the evaluation for one individual teacher, the inference can be reversed when the teacher is close to the threshold. As we see here, in total there are five teachers whose evaluation changes in different approaches and none of them shows very strong evidence of being ineffective in either approach, which illustrates the importance of quantifying the robustness of the inferences in evaluating individual teachers based on VAMs.

As mentioned before, the last two columns in Table 5 present the results of selective replacement approaches for the five teachers who are below the threshold in the DOLS approach. Specifically, the left column (DOLS_het_π1) applies the successive extreme replacement approach and the right column (DOLS_het_π2) applies the purposeful replacement approach. For example, for teacher A, the four lowest score students (33.33%) in the class must be replaced with average students to get this teacher to the threshold. Alternatively, 26.93% of below class average students must be replaced. The motivation for the selective replacement is to characterize how sensitive the teacher evaluation is to potential heterogeneous/outlier effects. That is, we want to use this information to capture not only how far the VAM (which is also the class average gain score) is from the threshold, but also to what extent the difference between VAM and threshold is due to students with very low gain scores in the class (i.e., tail part of the distribution). To illustrate how the selective replacement approaches achieve this goal, we present the results together with gain score distributions for the five teachers, as shown in Figure 2.9, where the red solid line represents the threshold and the black dashed line represents each teacher's VAM, which is also the class average gain score.

First, we compare teacher E and I. As shown in Figure 2.9, both E and I have students with very low gain scores. But teacher I's VAM is much closer to the threshold than E. Reflected in

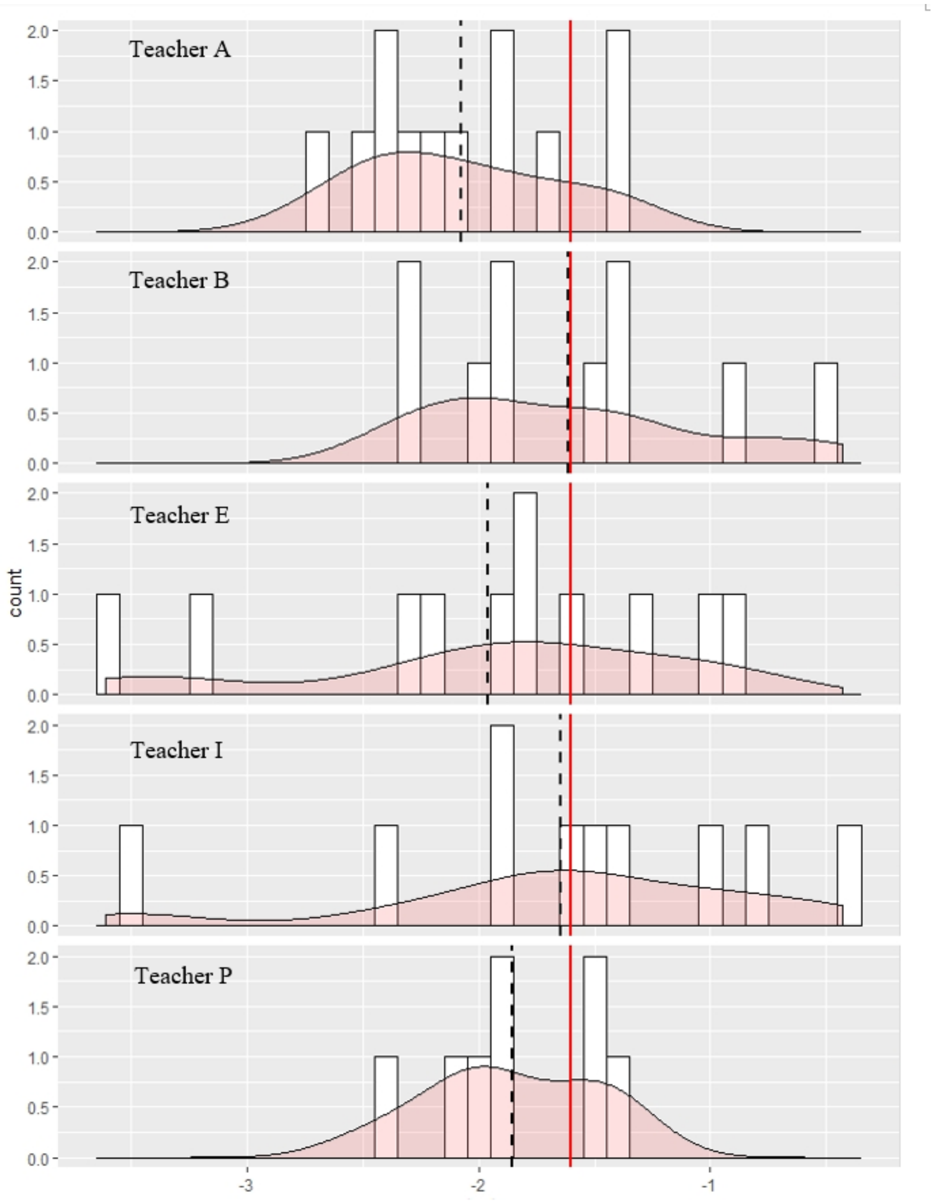


Figure 2.9: Gain score distributions for 5 teachers below the threshold in the DOLS approach.

Table 2.1, the selective replacement approach tells us that teacher I only needs to replace very few low-end students to get to threshold while teacher E needs to replace more low-end students. Second, we compare teacher E and P. E is a little further away from the threshold compared to P: 29.27% of E's students must be replaced with average students to change his/her evaluation and 22.67% of P's students must be replaced with average students to change his/her evaluation. But E has a few students with extremely low gain scores. As such, the selective replacement tells us that E's evaluation is more sensitive to students in the lower tail than teacher P.

2.7 Discussion

This Chapter discusses how the case replacement approach can be applied in the VAM context, accounting for spillover and heterogeneity. The general approach is very much like that in the MST context: to quantify the strength of evidence for an inference by considering how many observed cases need to be replaced with unobserved cases to change the inference (Frank et al., 2013). As we have shown, this is a very powerful non-parametric framework that can be easily generalized for violations of SUTVA and presence of heterogeneity. The extensions to account for spillover and heterogeneity are also essentially the same in both MST and VAM contexts.

Meanwhile, it is important to note two important differences in MST and VAM. First, sampling variability is accommodated by statistical significance attached to the threshold in the MST context. This is also the most general situation in empirical research. However, in the context of VAM, the threshold is arbitrary and thus we consider sampling variability with the number of cases that need to be replaced to invalidate the inference. In other words, instead of adding sampling variability to the threshold, we accommodate the sampling variability in the index of inference robustness. Second, the heterogeneity in VAM only exists within class (one level) but the MST has two levels of heterogeneity. Importantly, we argue that the second level of heterogeneity really needs more attention from researchers since the cross-site variation in treatment effects always has significant policy implications. The study of cross-site variation also leads to more research about how and why an intervention works in certain contexts but not others. We argue that our approach has a strong potential in helping researchers, as well as policymakers in understanding and interpreting these important variations.

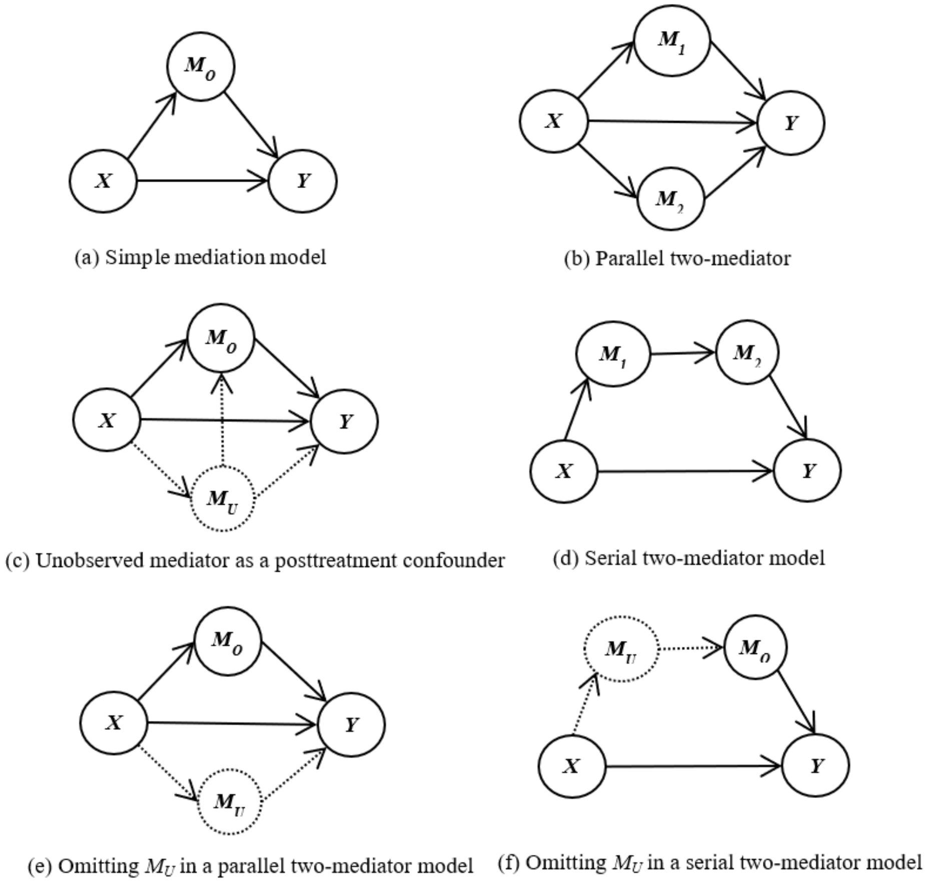
CHAPTER 3

UNOBSERVED MEDIATOR IN A SINGLE-MEDIATOR MODEL

3.1 Introduction

Chapters 1 and 2 focus on how researchers can present strength of evidence in educational research with an intuitive framework so that all stakeholders can evaluate the effectiveness of one intervention against potential costs to better inform policy choices. Now Chapters 3 will turn to the second goal stated in the introduction: exploring why an intervention works through mediation mechanisms so that policy choices may be better informed. Under a simple mediation model, a binary variable X is randomly assigned (e.g., treatment vs. control groups) and causes change in the outcome variable Y. A mediator variable M_O explains one mechanism or process through which X affects Y. In other words, the mediator is intermediate in the causal sequence that explains why the intervention (X) causes the outcome (Y) (e.g., Baron & Kenny, 1986; MacKinnon, 2008). This mediation process with a single mediator is presented in Figure 3.1(a), where the product of the effects from X to M_O and from M_O to Y is defined as the indirect effect via the mediator M_O (e.g., Hayes, 2013; MacKinnon, 2008).

The basic single mediator model can also be extended to more than one simultaneous mediator. Figure 3.1(b) presents a two-mediator parallel mediation model, where there are two simultaneous mediation processes, one through each mediator, connecting X to Y. Preacher and Hayes (2008) discussed approaches to test hypotheses for individual mediators and contrast the magnitude of indirect effects for different mediators in a multiple mediation model like this, where more than one simultaneous mediators are involved to explain why the intervention X causes the outcome Y. Researchers have emphasized the importance of testing multiple mediators in a single model rather than in separate models to prevent potential parameter bias due to omitted variables (e.g., Hayes, 2013; Judd & Kenny, 1981; Preacher & Hayes, 2008). In many cases, however, a single mediation model may be tested because a researcher has not measured another relevant potential mediator or



(a) Simple mediation model

(b) Parallel two-mediator

(c) Unobserved mediator as a posttreatment confounder

(d) Serial two-mediator model

(e) Omitting M_U in a parallel two-mediator model

(f) Omitting M_U in a serial two-mediator model

Figure 3.1: Simple mediation and dual mediator designs.

is unable to measure a theoretically relevant mediator. In such instances, the second potentially relevant mediator is unobserved (denoted by M_U). For example, suppose a researcher studying the impact of tracking/grouping students (X) on students' learning outcomes (Y) is interested in potential mediators that might explain associations between tracking and learning. If the researcher hypothesizes that stigma and teachers' mindset might both be mediators of interest but teachers' mindset was not measured under the research design, then in such an instance stigma is observed (M_O) and teachers' mindset is unobserved (M_U).

When M_U is associated with the observed mediator (denoted by M_O) (Figure 3.1(c)), it may threaten inference of the mediation effect via M_O. In this case, M_U is also a posttreatment confounder for the mediation path from X to M_O to Y as M_U explains the M_O to Y relation and is caused by X (e.g., Fritz, Kenny, & MacKinnon, 2016). Many recent studies have investigated influence of potential confounders on causal mediation inferences by providing approaches to sensitivity analyses (e.g., Hong, Qin, & Yang, 2018; Imai, Keele, & Tingley, 2010; Imai, Keele, & Yamamoto, 2010; Imai & Yamamoto, 2013; VanderWeele, 2015). Specifically, sensitivity analyses may describe the responsiveness of an estimated effect or the robustness of a causal inference to a potential confounder.

However, based on our knowledge, existing sensitivity analysis strategies are either 1) intended primarily for unobserved confounders, which are not potential mediators (e.g., Imai, Keele, & Yamamoto, 2010; Imai & Yamamoto, 2013; VanderWeele, 2010), or 2) use an alternative counterfactual framework targeting the “natural indirect effect” (NIE), “natural direct effect” (NDE) or “controlled direct effect” (CDE) (e.g., Hong et al., 2018; VanderWeele, 2015; VanderWeele & Chiba, 2014). Under models with multiple mediators, NIE, NDE and CDE can be very different from the specific indirect and direct effects defined from a path-specific perspective. For example, the NIE is defined as the difference in outcome Y if the mediator of interest changed to what it would have been if the exposure X changed to control (assuming X is binary for simplicity), while the exposure X stays at the treatment condition. In our dual mediator model of Figure 3.1(c), then the NIE transmitted through M_O includes not only the pathway $X \rightarrow M_O \rightarrow Y$ but also $X \rightarrow M_U \rightarrow M_O \rightarrow Y$. Additionally, the NDE is defined as the difference in outcome Y if only X changes from control to treatment but the mediator of interest M_O stays at the level that would have taken under the control condition of exposure X. Then in Figure 3.1(c), the NDE includes both $X \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$. Thus, the sensitivity approach under the counterfactual framework cannot be applied directly if we take the path-specific perspective commonly used in psychology (e.g., Hayes, 2013; MacKinnon, 2008).

Extending previous research, the goal of this study is to examine whether and how omitting an alternative mediator that is also a confounder can bias causal mediation effect estimates from the path-specific perspective. Furthermore, we propose a sensitivity analysis to evaluate robustness of a causal mediation inference to an unobserved mediator.

3.2 Dual-Mediator Designs

Hayes (2013) presented a complex serial mediation model with two mediators (Figure 3.1(c), with the dotted lines) where one mediator (M_U) has a serial or predictive effect on the second mediator (M_O), en route to Y. The parallel or serial mediation models (Figure 3.1(e) or 3.1(f)) are special cases of the complex mediation model presented in Figure 3.1(c). Specifically, when the path coefficient for $M_U \rightarrow M_O$ is zero, the model in Figure 3.1(c) becomes the parallel mediation model in Figure 3.1(e). When the path coefficients for both $X \rightarrow M_O$ and $M_U \rightarrow Y$ are zero, the model in Figure 3.1(c) becomes the serial mediation model in Figure 3.1(f). We focus on the complex serial mediation model (Figure 3.1(c), with the dotted line), as a general model, to evaluate whether and how an unobserved mediator may bias estimation of the direct and indirect effects via the observed mediator. Specifically, we study how excluding M_U affects estimation of the specific mediation effect via M_O, defined as the product of the paths from X to M_O and M_O to Y. That is, when M_U is no longer observed in our analysis, all the pathways that relate to this omitted mediator M_U are excluded from the analysis, as shown by dotted lines in Figure 3.1(c). To illustrate potential effects by omitting such a mediator, we will apply a real data example also used by Hayes (2013) about how media use affects behaviors.

3.3 Illustrative Data Example about Consequences of Omitting an Alternative Related Mediator

Illustrative data were originally drawn from an experimental study conducted by Tal-Or, Cohen, Tsfati, and Gunther (2010). The research was to test a hypothesis about the influence of media use: whether media (X) affects people's attitudes or behaviors (Y) through changing people's perceptions regarding how other people may be influenced by the media (M_O). For example, when a person Ashley reads a media report, she may tend to predict that others in her community will respond to this report in certain ways. This prediction can further affect Ashley's attitudes or behaviors. To test this hypothesis, participants were randomly assigned into two groups (X). Both groups were asked to read a newspaper article about an economic crisis that can affect the price and supply of sugar in Israel. However, one group were told that this article came from the front page of a major newspaper whereas the other group were told that the article appeared in the middle of an economic supplement of the newspaper. A condition variable (X, denoted by COND) was used to indicate whether the participant was manipulated to consider this article as a front-page article or interior-page article. After a participant finished reading the article, he/she was asked how much he/she believed that others in the community would be encouraged to buy sugar after they read this article. This presumed media influence served as one mediator (M_O, denoted by PMI) in the model. The participants were also asked about how important they thought this article was. This perceived issue importance was a second mediator (M_U, denoted by IMPORT) in the model. Finally, the participants in both groups were asked about how soon they intended to buy sugar and how much they would buy. These responses were then aggregated to generate a variable that measured their intention to buy sugar. This is then the outcome variable (Y, denoted by REACTION).

The first fitted model includes two mediators PMI and IMPORT as presented in the left panel of Figure 3.2. For the mediation path via PMI, it is hypothesized that others are more likely to be affected by an article appearing on the front page than on the interior page. Therefore, before others act to buy sugar forcing the price up, one would act as soon as possible to purchase sugar when it is available at an acceptable price. For the mediation path via IMPORT, people can infer

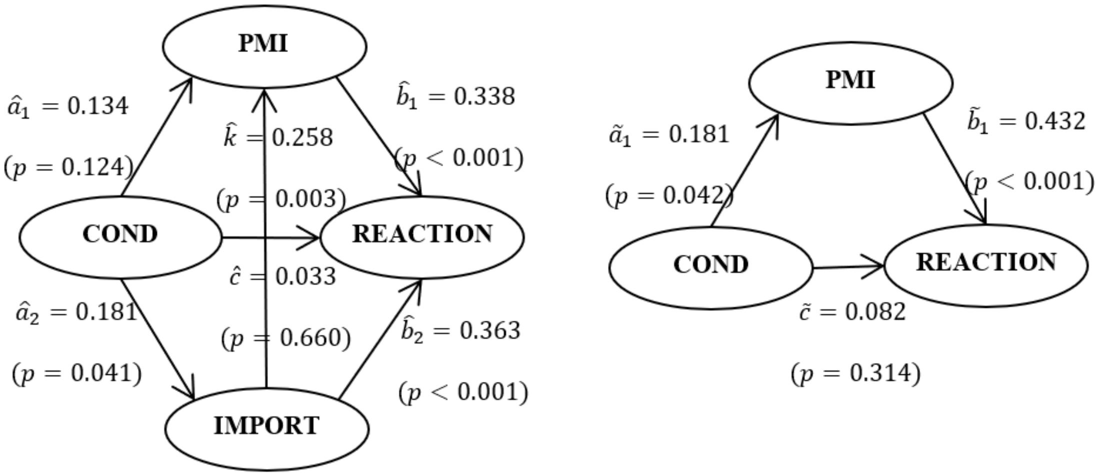


Figure 3.2: Illustrative data example of presumed media influence.

the importance of the article based on where it is published and consequently, people act upon the importance of the issue. Moreover, IMPORT is also presumed to predict media influence, under the hypothesis that the more important people believe the article is, the more likely they believe others would be influenced by that article.

We fit the model in the left panel of Figure 3.2 to data using standardized variables[1]. The estimated path coefficients were therefore standardized coefficients. Both the effects from COND to IMPORT and from IMPORT to REACTION are significantly positive. Using the joint significance approach (Fritz & MacKinnon, 2007), the estimated indirect effect through IMPORT was $0.181 \cdot 0.363 = 0.066$. The 95% percentile bootstrap confidence interval of the indirect effect is 0.001 to 0.150 (based on 5,000 bootstrap samples). In contrast, the other indirect effect via presumed media influence was not significant, with a 95% percentile bootstrap confidence interval of -0.013 to 0.114 (based on 5,000 bootstrap samples). That is to say, the specific mediated effect of the article's location did not have statistically significant influence on participants' reactions through presumed media influence. Additionally, the predictive relationship from IMPORT to PMI was significantly positive (0.258, $p = 0.003$). This confirms the hypothesis that as participants think the issue is more important, they have a stronger belief that others are going to be affected by that article.

Now we ask, "What would happen if the mediator IMPORT was excluded from the fitted model?" This may be the case if an alternative theoretically important mediator (such as IMPORT) is not measured in the research design, as can occur in almost any example of a mediation analysis. In this case, PMI is the only observed mediating variable included in the model. The results are presented in the right panel of Figure 3.2. After excluding IMPORT, the specific indirect effect via PMI changed from being non-significant to significantly positive: both paths in the indirect effect are statistically significantly positive and the indirect effect has a 95% confidence interval of 0.078 to 0.042 (based on 5,000 bootstrap samples). Specifically, the direct path from article location to presumed media influence increased from 0.134 to 0.181 and the direct path from presumed

[1] We standardized all variables for analysis to be consistent with our later derivation. Therefore, the coefficients are different from those in Hayes (2013).

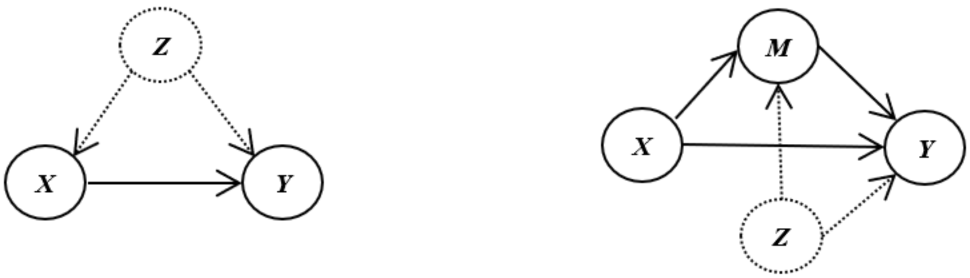
media influence to participants' reaction increased from 0.338 to 0.432. Therefore, excluding IMPORT, both paths for this indirect effect via presumed media influence were greater. Omitting this alternative mediator has given us a different conclusion that the indirect effect via presumed media influence is significantly positive. We also observe that in both models the direct path from article location to participants' reaction was not statistically significant from 0 but the point estimate increases from 0.033 to 0.082 once we exclude the IMPORT mediator from our model.

It is important to note that when we compare the two fitted models above, we are mainly focusing on two estimates: 1) the specific indirect effect via PMI, defined as the product of the paths from COND to PMI and from PMI to REACTION; 2) the direct path from COND to REACTION. The specific indirect effect defined as the product of two path coefficients (e.g., Hayes, 2013) is not equal to either the natural indirect effect or the randomized interventional analogue of the natural direct effect (e.g., VanderWeele, 2015). The direct effect we are interested here is also different from both natural and control direct effect defined in the counterfactual framework (e.g., VanderWeele & Chiba, 2014).

3.4 Unobserved Causally Related Mediator as Posttreatment Confounder

The example above illustrates that omitting an alternative mediator can alter estimates of the mediating path coefficients for PMI and change the inference for this mediating effect. To understand why omitting IMPORT can generate such effects, we first present a general representation of a confounder in Figure 3.3(a): a confounder Z of X and Y can bias the estimate and inference of the effect from X to Y by serving as an alternative cause of both X and Y. Following the illustrative data example, in such a case IMPORT is associated with both the mediator PMI and the outcome REACTION. Thus, if IMPORT is omitted, it serves as a potential confounder for PMI and REACTION.

Previous studies on sensitivity analysis in the mediation literature have focused on the case where a potential confounder (e.g., IMPORT) between a mediator (e.g., PMI) and an outcome (e.g., REACTION) is assumed to be independent from the input variable (e.g., COND) (Figure 3.3(b)).



(a) Confounder Z of the X to Y relation (b) Confounder Z of the mediator-outcome

Figure 3.3: Confounder in mediation.

For example, Imai, Keele, and Yamamoto (2010) presented a sensitivity analysis based on residual correlations to deal with unmeasured pre-treatment covariates "that confound the relationship between the mediator and the outcome". The "pre-treatment" covariates precede the input variable X and therefore are not influenced by X. Fritz et al. (2016) combined the effects of measurement errors and omitting confounders on bias of the mediation effect, but they considered the confounder Z as independent of X throughout their analytical study (Figure 3.3(b)).

Meanwhile, empirical studies across the social sciences have shown considerable evidence for the existence of multiple causal mechanisms that may involve simultaneous or causally related mediators. For example, Bekman, Cummins, and Brown (2010) examined a parallel multiple-mediator model where depression affected adolescent alcohol use through simultaneous mediators perceptions and expectancies. Imai and Yamamoto (2013) discussed several empirical studies where both content and importance mechanisms mediated the effect of political issue framing on citizens' political opinion and behaviors. Singh, Chen, and Wegener (2014) showed evidence for several sequential multiple-mediator models for how attitude similarity affected inferred attraction through several mediators. As such, the parallel mediation (Figure 3.1(b)) and serial mediation (Figure 3.1(d)) models are often observed to characterize mediational effects, which emphasizes the importance of understanding unobserved/omitted mediators in cases where there is an association between the predictor and the unobserved mediator.

Therefore, it is of critical importance to fill the research gap regarding how unobserved/omitted mediators affect estimates and inferences for observed mediators as these post-treatment confounders can easily occur in practice. In this chapter, we focus on the two-mediator model discussed by Hayes (2013) (Figure 3.1(c) with dotted lines) to study how an unobserved/omitted mediator (M_U) may bias direct and indirect effect estimates via the observed mediator (M_O). This selected two-mediator model also makes our study results generalizable from two perspectives. First, as discussed before, this model can be simplified to other popular two-mediator designs: parallel two-mediator model (Figure 3.1(e)) and serial two-mediator model (Figure 3.1(f)). Second, by setting the path coefficient from X to the omitted mediator M_U to zero, the studied scenario becomes the same as that for previous studies of the sensitivity of mediation effects (e.g., Fritz et al., 2016; Imai, Keele, & Yamamoto, 2010; Imai & Yamamoto, 2013; VanderWeele, 2010). Therefore, we can also compare our findings with these previous studies and extend their findings.

3.5 Goals of the Study

This study has three goals. Our first goal is to establish conditions under which omitting a mediator can yield consistent results. However, we show that the conditions to obtain consistent estimates for the direct and indirect effect via M_O are not the same, so that one cannot simultaneously obtain consistent direct and indirect effects. Second, recognizing that these conditions may be difficult to meet in practice, we evaluate how omitting M_U biases the mediation effect estimate via M_O. For example, we find that when the path coefficients related to M_U are all positive or all negative, the specific indirect effect via M_O is always overestimated while the direct effect from X to Y can be either positively or negatively biased. We further show that as the path coefficient of $M_U \rightarrow M_O$ becomes larger, the positive bias in estimating the indirect effect via M_O becomes larger, but the estimation of the direct effect $X \rightarrow Y$ changes from overestimation to underestimation. Third, we also seek to identify the magnitude of bias under different scenarios in order to contextualize the potential threats to inference by omitting the post-treatment confounder M_U. Finally, stemming from the first three goals, we seek to propose a sensitivity analysis to assessing the robustness of the causal mediation inference to omission of an unobserved mediator that is confounded with the M_O paths through its relationship to X and Y.

We present our analytical findings in terms of path coefficients as well as correlations among variables in the underlying model. The path coefficient framework shows how different levels of parameters for causal pathways in the true model affect the direction and magnitude of inconsistency. Alternatively, the correlation framework shows how estimates of path coefficients vary with different magnitudes of correlations among the variables. We have elected to present both frameworks because the parameter framework allows us to understand how inconsistency is generated when we omit the other mediator by assuming we know the "truth", while the correlation framework facilitates our development of sensitivity analysis methods that can be directly applied by substantive researchers in real studies. We expect that these two frameworks can complement each other to fulfill a more comprehensive picture so that a solid foundation can be built for substantive researchers to deal with an unobserved mediator as a potential confounder. Throughout, we will draw on the previously discussed illustrative example of presumed media influence to demonstrate the effect of omitting a related mediator.

3.6 Inconsistency When M_U is Omitted

The first and primary goal of this paper is to derive the effect estimates with an omitted confounding mediator and their deviations from the true effects in the underlying model. Specifically, we would like to study the differences between $\tilde{a}_1, \tilde{b}_1$, $\tilde{c}$ and a_1, b_1, c, as shown in Figure 3.4.

We applied the Law of Iterated Expectation to derive $\tilde{a}_1, \tilde{b}_1, \tilde{c}$. The Law of Iterated Expectation states that $E(Y|X) = E\ [E(Y|X,Z)|X]$ where X, Y and Z are three variables and $E(Y|X)$ represents the conditional expectation (or conditional mean) of Y given X (Wooldridge, 2009). The Appendix has shown the derivation details and the results are presented in Equations 3.1 through 3.4 as

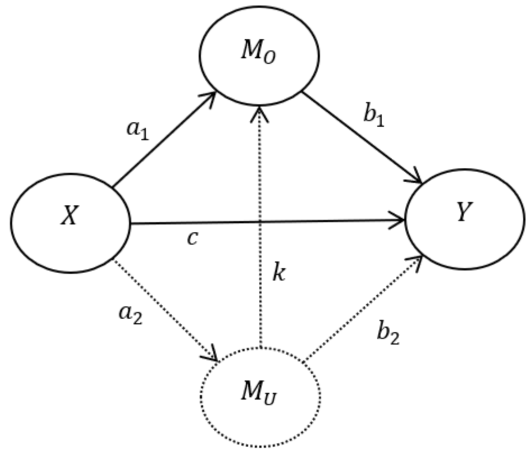
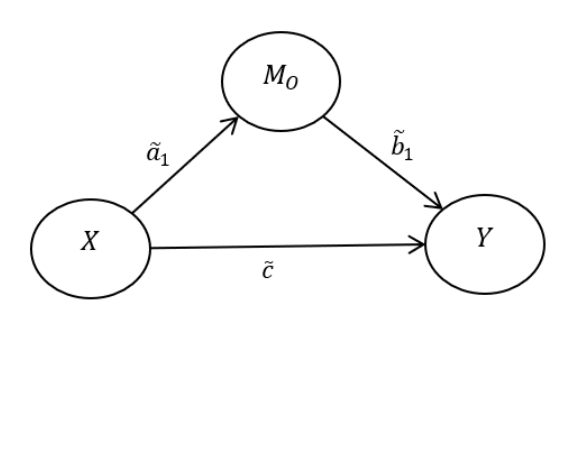
True Model	Model if omitting M_U
	
$M_{Oi} = k \cdot M_{Ui} + a_1 \cdot X_i + \varepsilon_{M_{Oi}}$ $M_{Ui} = a_2 \cdot X_i + \varepsilon_{M_{Ui}}$ $Y_i = b_1 \cdot M_{Oi} + b_2 \cdot M_{Ui} + c \cdot X_i + \varepsilon_{Y_i}$	$M_{Oi} = \tilde{a}_1 \cdot X_i + \varepsilon_{M_{Oi}}$ $Y_i = \tilde{b}_1 \cdot M_{Oi} + \tilde{c} \cdot X_i + \varepsilon_{Y_i}$

Figure 3.4: True model with two mediators and the model omitting M_U.

follows.

$$\tilde{a}_1 = a_1 + k \cdot a_2, \tag{3.1}$$

$$\tilde{b}_1 = b_1 + b_2 \cdot k \cdot \frac{1 - a_2^2}{1 - (k \cdot a_2 + a_1)^2}, \tag{3.2}$$

$$\tilde{c} = c + b_2 \cdot \frac{a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)}{1 - (k \cdot a_2 + a_1)^2} = c + b_2 \cdot \frac{\rho_{X,M_U} - \rho_{M_O,M_U} \cdot \rho_{X,M_O}}{1 - (k \cdot a_2 + a_1)^2} \tag{3.3}$$

$$\tilde{a}_1 \tilde{b}_1 = a_1 b_1 + k \cdot a_1 \cdot b_1 + \frac{k \cdot (a_1 + k \cdot a_2)(1 - a_2^2)}{1 - (k \cdot a_2 + a_1)^2} \tag{3.4}$$

As such we obtain $\tilde{a}_1, \tilde{b}_1$, and $\tilde{c}$ as functions of true parameters. Specifically, a_1 and b_1 represent the true effects via the observed mediator M_O, while a_2 and b_2 represent the true effects via the omitted mediator M_U. In our illustrative example, the product of a_1 and b_1 is the true mediation effect through the mediator PMI, while the product of a_2 and b_2 is the true mediation effect through the omitted mediator IMPORT. k stands for the path coefficient from IMPORT to PMI. As shown by Equations 3.1 through 3.4, only under very stringent conditions are $\tilde{a}_1, \tilde{b}_1$, and $\tilde{c}$ equal to a_1, b_1, and c. In the following subsections, we will show these stringent conditions, and how the key parameters influence the differences between $\tilde{a}_1$, $\tilde{b}_1$, $\tilde{c}$ and the true parameters a_1, b_1,

c.

To note, the derivation here is at the population level and therefore is free of sampling error. Technically speaking, the difference between $\tilde{a}_1, \tilde{b}_1, \tilde{c}$ and a_1, b_1, c are *inconsistencies*, not *bias*. That is, $\tilde{a}_1, \tilde{b}_1, \tilde{c}$ are the estimates we get even when we have all population data available (i.e., sample size $n \to \infty$). In all following discussion, "bias" in the population parameters is a rough way to describe "inconsistency" for an audience more comfortable with conceptualizations of bias.[2]

3.6.1 Conditions for consistent estimates when omitting M_U

Next, we examine conditions under which omitting M_U can result in consistent results. If an empirical researcher can justify their studies as meeting these conditions, there is less concern about omitting M_U.

From Equation 3.1, $\tilde{a}_1 = a_1$ if and only if $k = 0$ or $a_2 = 0$. When $k = 0$, there is no causal relation between the two mediators, which reduces to the parallel mediation model represented in Figure 3.1(e). In our illustrative data example, this condition is satisfied when IMPORT has no effect on PMI. Under this condition, excluding IMPORT does not affect the effect estimate from COND to PMI. Alternatively, $\tilde{a}_1 = a_1$ when $a_2 = 0$, or there is no effect of the treatment variable X on the omitted variable M_U. This condition is equivalent to the situation that the omitted M_U is only a confounder of the $M_O - Y$ relation and is not caused by X.

Based on Equation 3.2, to satisfy $\tilde{b}_1 = b_1$, we can have (1) $b_2 = 0$, indicating no effect of M_U on Y, (2) $k = 0$, indicating no effect of M_U on M_O (a parallel two-mediator model), or (3) $a_2^2 = 1$, indicating the effect of X on M_U is -1 or 1. If condition (3) is true, then the omitted mediator M_U

[2] Take the $\tilde{a}_1$ as an example to see why we are deriving *inconsistencies* rather than *bias*. The derivation in the Appendix shows that a_2 comes from β_1 where $E(M_U|X) = \beta_1 \cdot X$. For one sample, we can express $\hat{\beta}_1$ as $(X'X)^{-1}(X'M_U)$ where X and M_U are sample data for X and M_U. This is a nonlinear function of X. If we really want to derive the magnitude of bias, we need to know $E((X'X)^{-1}(X'M_U))$. Since only probability limit goes through nonlinear function but expectation cannot, we can only know $plim((X'X)^{-1}(X'M_U))$ but not $E((X'X)^{-1}(X'M_U))$. That is why in this chapter (also chapter 4) I am in fact deriving *inconsistency* which is at the population level.

is fully determined by the treatment X[3]. In our illustrative example, condition (3) indicates that the value of IMPORT is solely dependent on COND.

In terms of the direct effect $\tilde{c}$, based on Equation 3.3, $\tilde{c} = c$ if and only if $a_2 = (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)$ or $b_2 = 0$. The first condition $a_2 = (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)$ is equivalent to $\rho_{X,M_U} = \rho_{M_O,M_U} \cdot \rho_{X,M_O}$, where $\rho_{X,M_U}, \rho_{M_O,M_U}, \rho_{X,M_O}$ are correlations between X and M_U, between M_O and M_U, and between X and M_O, respectively. In the illustrative example, this means the correlation between COND and the omitted mediator IMPORT is equivalent to the product correlation between PMI and IMPORT × the correlation between COND and PMI. The second condition $b_2 = 0$ indicates no effect of M_U on Y.

In sum, as long as the mediation effect via the omitted mediator $M_U$is nonzero ($a_2 \neq 0$ and $b_2 \neq 0$) and M_U has a nonzero effect on the observed mediator M_O ($k \neq 0$), the estimates for a_1, b_1, c will be inconsistent (i.e., $a_1 \neq \tilde{a}_1, b_1 \neq \tilde{b}_1, c \neq \tilde{c}$). Furthermore, the conditions to obtain consistent direct and indirect effects are not the same, so that one cannot simultaneously obtain accurate direct and indirect effects. An unbiased indirect effect can be obtained when M_U is omitted when M_U has a zero effect on the observed mediator M_O, or the underlying model is a parallel two-mediator model. In fact, under a parallel two-mediator model, not only can we obtain an consistent indirect effect, but also consistent estimates of a_1 and b_1. A consistent direct effect from X to Y can be obtained when M_U is omitted when $\rho_{X,M_U} = \rho_{M_O,M_U} \cdot \rho_{X,M_O}$ or $b_2 = 0$.

3.6.2 Direction of inconsistency when omitting M_U

From the previous discussion, we note that conditions for obtaining unbiased estimates of direct and/or indirect effects are difficult to justify in practice. This leads us to the following discussion to achieve our second goal: to examine the direction and magnitude of bias of direct and indirect effect estimates when omitting M_U.

From Equation 3.1, $\tilde{a}_1 > a_1$ as long as k and a_2 are in the same direction. That is, when

[3] We standardized all variables in our derivation. Therefore, this condition of $a_2^2 = 1$ is equivalent to zero error term, indicating the omitted mediator only depends on the treatment variable.

k and a_2 are both positive or negative, a_1 is overestimated. Similarly, $\tilde{b}_1 > b_1$ when k and b_2 are in the same direction (either both positive or both negative, but not zero, see Appendix for detailed proof). Therefore, the indirect effect estimate via M_O is overestimated when k, a_2 and b_2, the three M_U-related path coefficients, are all positive or all negative. The amount of bias is $k \cdot a_2 \cdot b_1 + \frac{k \cdot b_2 \cdot (a_1 + k \cdot a_2) \cdot (1 - a_2^2)}{1 - (k \cdot a_2 + a_1)^2}$. In our earlier example, the three path coefficients related to IMPORT were all positive. Correspondingly, when IMPORT was excluded from the model, the indirect effect through PMI became larger, from 0.045 to 0.078.

In terms of the direct effect c, the direction of bias depends on whether b_2 and $[a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)]$ have the same sign. In the Appendix, we show that $[a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)]$ is equivalent to $(\rho_{X,M_U} - \rho_{X,M_O} \cdot \rho_{M_O,M_U})$, which is also the numerator of the partial correlation between X and M_U conditional on M_O. When $b_2 > 0$ and $\rho_{X,M_U} > \rho_{X,M_O} \cdot \rho_{M_O,M_U}$, c is overestimated. In contrast, when $b_2 > 0$ and $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$, c is underestimated. One way to think about the sign of $(\rho_{X,M_U} - \rho_{X,M_O} \cdot \rho_{M_O,M_U})$ is to consider the observed mediator M_O as a common outcome of the omitted variable M_U and the treatment variable X. When we have $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$, then ρ_{X,M_U} changes its sign once paritialling out the M_O. In this case, M_O functions similarly to a suppressor since it distorts the relation between M_U and X (e.g., Cohen & Cohen, 1983; Rosenberg, 1968).

In our illustrative example of presumed media influence, the path coefficient from PMI to the outcome REACTION (b_2) was positive. Additionally, the correlation between COND and IMPORT was larger than the product of correlations between COND and PMI and between PMI and IMPORT: $\rho_{X,M_U} > \rho_{X,M_O} \cdot \rho_{M_O,M_U}$. Accordingly, the direct effect became larger, from 0.033 to 0.082, when IMPORT was excluded from the model.

3.6.3 How inconsistency changes with M_U-related parameters.

Next we will demonstrate how the omission of M_U can generate bias with regards to M_O under different scenarios. To achieve this goal, we will manipulate each of the parameters related to M_U. The discussion here will exhaust all possible scenarios about how each M_U-related path coefficient affects bias regarding M_O, in terms of increasing or decreasing the bias. For example, the path coefficient from M_U to M_O (k) is always positively associated with the bias in estimating b_1, no matter what values other parameters take on. In contrast, the path coefficient from X to M_U (a_2) can be either positively or negatively related to the bias in $\tilde{b}_1$ depending on values of other parameters. To simplify, we constrain our discussion to all parameters in the true model being positve[4].

We start our discussion with the effect from M_U to M_O (represented as k). Figure 3.5 plots the bias in estimating a_1, b_1, c against different values of k, while all the other parameters are fixed at certain values: $a_1 = a_2 = 0.2$, $b_2 = 0.1$. First the magnitude of bias in estimating both $X \rightarrow M_O(a_1)$ and $M_O \rightarrow Y(b_1)$ increases as path k becomes larger. Focusing on the effects $X \rightarrow M_U \rightarrow M_O$ in the true model (left panel in Figure 3.4), the larger the k the more some of the effect attributed to M_O should be attributed to M_U as a mediator. Therefore, we overestimate a_1 and the difference between $\tilde{a}_1$ and a_1 grows as k gets bigger. Similary, k plays a role in the chain of effects $X \rightarrow M_U \rightarrow M_O \rightarrow Y$. Here, the larger the magnitude of k the larger the bias in the estimated effect $M_O \rightarrow Y$. Thus, as k increases the bias in both $\tilde{a}_1$ and $\tilde{b}_1$ increase, leading to more serious overestimation of the indirect effect $X \rightarrow M_O \rightarrow Y$.

Figure 3.5 also shows that after the positive bias in the estimated direct effect $X \rightarrow Y$ falls to 0 it continues decreasing to negative values as k increases to 1. That is, when k is relatively small we overestimate c while when k becomes larger we underestimate c. As k gets closer to 1, the magnitude of negative bias in $\tilde{c}$ keeps growing. This is consistent with our previous discussion about the direction of bias in $\tilde{c}$. As k becomes larger, the "suppression" function from the common outcome M_O becomes more serious as the difference between ρ_{X,M_U} and $\rho_{X,M_O} \cdot \rho_{M_O,M_U}$ becomes larger.

To summarize, when we manipulate k to vary from 0 to 1, we always overestimate the indirect effect via M_O and the bias becomes larger. However, the increase in k first countermands the overestimation in c and as k gets closer to 1 it finally leads to underestimation of the direct effect $X \rightarrow Y$. The Appendix also proves that this pattern of associations between k and the bias of $\tilde{a}_1$,

[4] See the Appendix for more detailed discussion about how we calculated these different scenarios by studying the first derivatives of the bias of $\tilde{a}_1$, $\tilde{b}_1$ and $\tilde{c}$ as a function of M_U-related parameters.

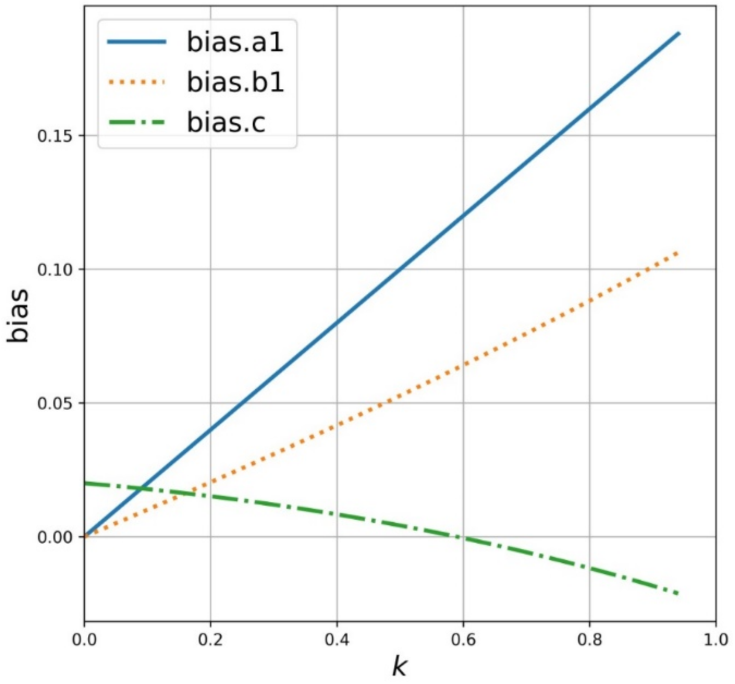


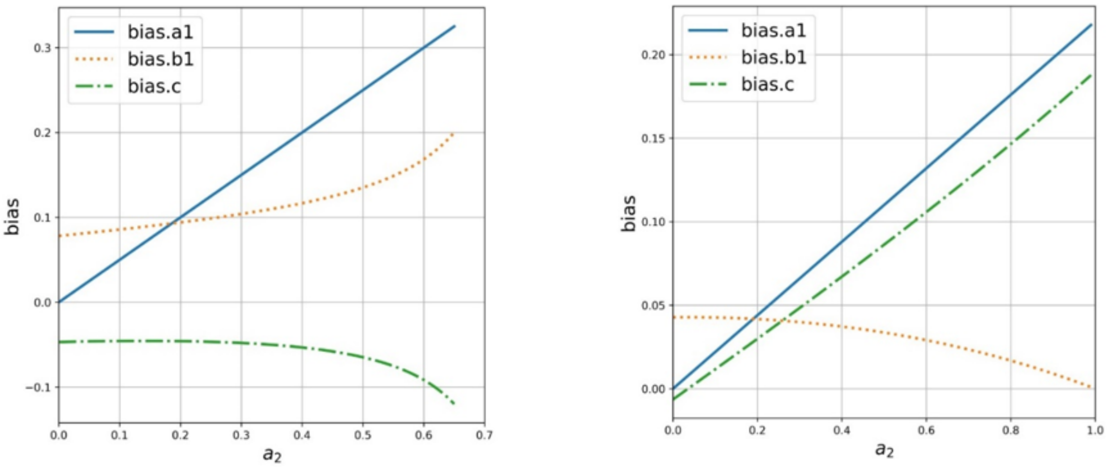
Figure 3.5: How bias changes with different levels of k.

$\tilde{b}_1$ and $\tilde{c}$ is consistent no matter what values other parameters take on.

Now we manipulate a_2 and see how this direct effect $X \rightarrow M_U$ relates to changes in bias. The results are presented in Figure 3.6(a) and Figure 3.6(b). In both figures, a larger direct effect $X \rightarrow M_U$ (a_2) is associated with larger bias in the estimated direct effect $X \rightarrow M_O$. The intuition is similar to how k affects the bias of $\tilde{a}_1$: the indirect effect $X \rightarrow M_U \rightarrow M_O$ is allocated to the biased effect $X \rightarrow M_O$ (reflected as $\tilde{a}_1$), where the omitted M_U plays a role of mediator. Consistent with our previous discussion, Figure 3.6(a) and 3.6(b) shows the direct effect $M_O \rightarrow Y$ is always overestimated (the bias is positive) (when all parameters are positive) while $\tilde{c}$ can be either positively or negatively biased.

Note that Figure 3.6(a) and 3.6(b) present different trends of changes in the bias of $\tilde{b}_1$ and $\tilde{c}$ as a_2 increases[5]. Figure 3.6(a) shows the bias of $\tilde{b}_1$ decreases until 0 as a_2 increases while Figure 3.6(b) shows the bias of $\tilde{b}_1$ increases as a_2 gets bigger. On the contrary, the bias of $\tilde{c}$ becomes

[5] The Appendix shows how we calculated these two different patterns by studying the partial derivatives of the bias in estimating a_1, b_1 and c with respect to a_2.



(a) How a_2 affects bias ($a_1 = 0.15, b_2 = 0.19, k = 0.22$).

(b) How a_2 affects bias ($a_1 = 0.6, b_2 = 0.1, k = 0.5$).

Figure 3.6: How bias changes with different level of a_2.

larger in Figure 3.6(a) but Figure 3.6(b) gives a decreasing curve for the bias in estimating c. The decreasing curve indicates the underestimation of c becomes more extreme as a_2 increases. The distinction between Figure 3.6(a) and 3.6(b) illustrates that the effect of a_2 on the bias in estimating b_2 and c relies on the values of other parameters. Specifically, Figure 3.6(a) presents the scenario where the magnitude of a_1 and k are both relatively small (with values of 0.15 and 0.22) while in Figure 3.6(b) a_1 and k are both relatively large (with values of 0.6 and 0.5).

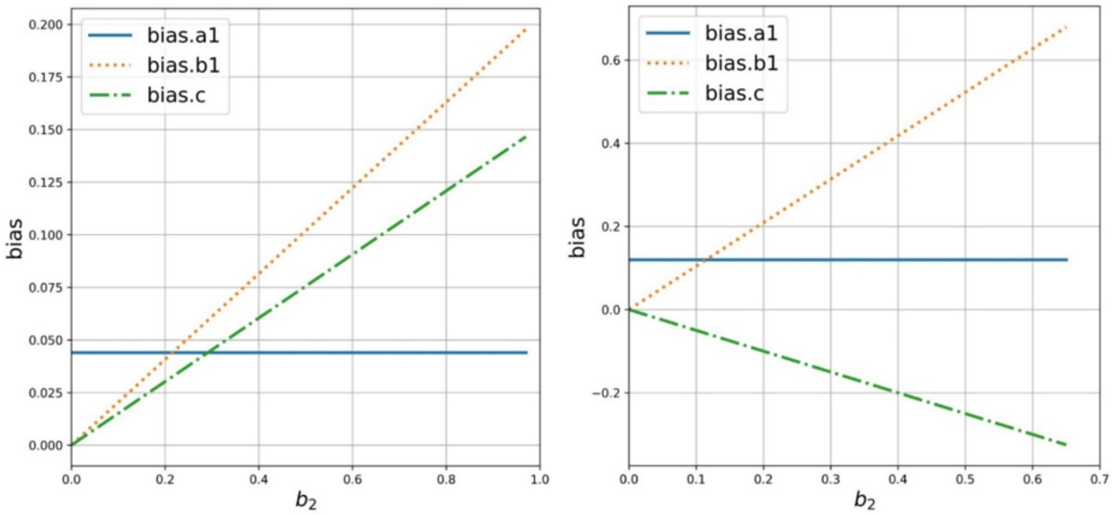
The last parameter that relates to M_U is the effect $M_U \rightarrow Y$, represented as b_2. Figure 3.7(a)[6] and Figure 3.7(b) [7] present how the bias in estimating a_1, b_1 and c vary with differing levels of b_2 under two different scenarios [8]. In both Figure 3.7(a) and 3.7(b), $\tilde{a}_1$ and $\tilde{b}_1$ are always positively biased. But the positive bias of $\tilde{b}_1$ becomes larger as b_2 gets closer to 1 while the level of b_2 has no effect on the bias in estimating a_1.

To see why b_2 has no effect on the magnitude of the bias of $\tilde{a}_1$, we focus on $X \rightarrow M_U \rightarrow M_O$ in the true model because this is where the bias of $\tilde{a}_1$ comes from (based on Equation 3.1). It is

[6] $a_1 = 0.2$, $a_2 = 0.2$, $k = 0.2$.

[7] $a_1 = 0.55$, $a_2 = 0.2$, $k = 0.6$.

[8] The Appendix shows how we calculated these two different patterns by studying the partial derivatives of the bias in estimating a_1, b_1 and c with respect to b_2.



(a) How b_2 affects bias ($a_1 = k = a_2 = 0.2$).

(b) How b_2 affects bias ($a_1 = 0.55, k = 0.6, a_2 = 0.2$).

Figure 3.7: How bias changes with different levels of b_2.

obvious then that b_2 plays no role in $X \longrightarrow M_U \longrightarrow M_O$. That is, the direct effect $M_U \longrightarrow Y$ (b_2) is not directly connected with the direct effect $X \longrightarrow M_O$ (a_1). Alternatively, we can consider this "no effect" by noticing that the path b_2 is invoked after the direct effect $X \longrightarrow M_O$ (a_1) in the sequence of causality. Equation 3.2 also shows that the bias in estimating b_1 is a linear function of b_2 as in both Figure 3.7(a) and 3.7(b).

The distinction between Figure 3.7(a) and 3.7(b) is how different levels of b_2 affect the bias of $\tilde{c}$. In Figure 3.7(a), c becomes overestimated and the positive bias grows but in Figure 3.7(b) we always underestimate c and more importantly, the negative bias becomes more serious as b_2 increases. This distinction is due to the direction of $\tilde{c}$'s bias while the effect of b_2 on the magnitude of $\tilde{c}$'s bias follows the same pattern in Figure 3.7(a) and 3.7(b). In Figure 3.7(a) we have $a_1 = k = a_2 = 0.2$ but in Figure 3.7(b), $a_1(= 0.55)$ and $k(= 0.6)$ are much larger than $a_2(= 0.2)$. Correspondingly, $\rho_{X,M_U} > \rho_{X,M_O} \cdot \rho_{M_O,M_U}$ for Figure 3.7(a) and $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$ for Figure 3.7(b). Based on the previous discussion, these two scenarios exemplify the cases in which c gets overestimated and underestimated, respectively. Importantly, in both scenarios, b_2 has no effect on the direction of $\tilde{c}$'s bias but only exhibits a positive influence on the magnitude of

$\tilde{c}$'s bias. When $\tilde{c}$ is positively biased, this influence then manifests as more serious overestimation. Alternatively, when the bias is negative, it results in more serious underestimation.

3.7 How Serious Inconsistency Could be at Different Levels of M_U-related Parameters

In the previous section we demonstrate how differing levels of M_U-related parameters may affect bias. In this section, we will solve for bias under different conditions of a_2, b_2 and k as small (0.1), medium (0.25) and large (0.4) to demonstrate how serious the bias could be and how large the bias is relative to the true effect. Starting from general situations, we will also discuss special situations where parallel ($k = 0$) or serial ($a_1 = b_2 = 0$) two-mediator model is the true underlying mediation process. To link our analysis with previous literature, we will also present situations where the omitted M_U is only a confounder for the mediator-to-outcome relation but not a mediator itself ($a_2 = 0$).

Table 3.1 depicts situations where the dual mediator design in Figure 1.7(c) (with dotted lines) is the true model, with all pathway coefficients being positive. Specifically, in all models, $a_1 = 0.2$, $b_1 = 0.15$ and $c = 0.1$ (except in Table 3.3 of serial mediation models where $a_1 = 0$). That is, the true indirect effect is 0.03 and the true direct effect is 0.1. Each row displays the bias in one scenario with different values of a_2, b_2 and k either small (0.1), medium (0.25) or large (0.4). Consistent with previous discussion, all cases in Table 3.1 generate positively biased values for the indirect effect via M_O from X to Y, since all M_U-related path coefficients (a_2, b_2 and k) take on positive values. But the *magnitude* of bias for the estimated indirect effect via M_O can vary substantially in different cases: in the worst scenario, large a_2, medium b_2 and large k generates a biased indirect effect around 0.09, almost three times as large as the true value of 0.03; while in the best scenario, small a_2, b_2 and k only generates an indirect effect of 0.034, just slightly larger than the true value of 0.03. The magnitude of bias for the estimated direct effect $\tilde{c}$ also varies greatly from case to case, with the worst scenario displaying a bias as large as 0.15. In our illustrative example, when IMPORT was included in the model, a_2 was a small to medium value of 0.181, k was a medium value of 0.258, and b_2 was a medium to large value of 0.363. As such, when

IMPORT was excluded, the indirect effect via PMI increased around 73.33%, from 0.045 to 0.078. The direct effect from COND to REACTION increased almost 150%, from 0.033 to 0.082.

Table 3.1: Bias in the Estimated Direct Effect of X on Y and Estimated Indirect Effect via M_O.

Parameters[a]			Bias				Bias/True value			
a_2	b_2	k	$(\tilde{a}_1 - a_1)$	$(\tilde{b}_1 - b_1)$	$(\tilde{c} - c)$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1)$	$(\tilde{a}_1 - a_1) / a_1$	$(\tilde{b}_1 - b_1) / b_1$	$(\tilde{c} - c) / c$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1) / a_1 b_1$
0.4	0.4	0.1	0.04	0.04	0.15	0.01	20.0%	23.8%	151.4%	48.5%
0.4	0.4	0.25	0.10	0.09	0.13	0.04	50.0%	61.5%	132.3%	142.3%
0.4	0.4	0.4	0.16	0.15	0.10	0.08	80.0%	102.9%	104.4%	265.3%
0.4	0.25	0.1	0.04	0.02	0.09	0.01	20.0%	14.9%	94.7%	37.8%
0.4	0.25	0.25	0.10	0.06	0.08	0.03	50.0%	38.5%	82.7%	107.7%
0.4	0.25	0.4	0.16	0.10	0.07	0.06	80.0%	64.3%	65.3%	195.8%
0.4	0.1	0.1	0.04	0.01	0.04	0.01	20.0%	5.9%	37.9%	27.1%
0.4	0.1	0.25	0.10	0.02	0.03	0.02	50.0%	15.4%	33.1%	73.1%
0.4	0.1	0.4	0.16	0.04	0.03	0.04	80.0%	25.7%	26.1%	126.3%
0.25	0.4	0.1	0.03	0.04	0.09	0.01	12.5%	26.3%	91.1%	42.1%
0.25	0.4	0.25	0.06	0.10	0.07	0.04	31.3%	67.1%	73.6%	119.4%
0.25	0.4	0.4	0.10	0.16	0.05	0.06	50.0%	109.9%	50.5%	214.8%
0.25	0.25	0.1	0.03	0.02	0.06	0.01	12.5%	16.5%	56.9%	31.0%
0.25	0.25	0.25	0.06	0.06	0.05	0.03	31.3%	42.0%	46.0%	86.3%
0.25	0.25	0.40	0.10	0.10	0.03	0.05	50.0%	68.7%	31.6%	153.0%
0.25	0.1	0.1	0.03	0.01	0.02	0.01	12.5%	6.6%	22.8%	19.9%
0.25	0.1	0.25	0.06	0.03	0.02	0.02	31.3%	16.8%	18.4%	53.3%
0.25	0.1	0.4	0.10	0.04	0.01	0.03	50.0%	27.5%	12.6%	91.2%
0.1	0.4	0.1	0.01	0.04	0.03	0.01	5.0%	27.6%	31.3%	34.0%
0.1	0.4	0.25	0.03	0.10	0.02	0.03	12.5%	69.5%	16.5%	90.7%
0.1	0.4	0.4	0.04	0.17	0.00	0.05	20.0%	112.1%	-0.3%	154.5%
0.1	0.25	0.1	0.01	0.03	0.02	0.01	5.0%	17.3%	19.6%	23.1%
0.1	0.25	0.25	0.03	0.07	0.01	0.02	12.5%	43.4%	10.3%	61.4%
0.1	0.25	0.4	0.04	0.11	-0.0002	0.03	20.0%	70.0%	-0.2%	104.0%
0.1	0.1	0.1	0.01	0.01	0.01	0.004	5.0%	6.9%	7.8%	12.2%
0.1	0.1	0.25	0.03	0.03	0.004	0.01	12.5%	17.4%	4.1%	32.1%
0.1	0.1	0.4	0.04	0.04	-0.0001	0.02	20.0%	28.0%	-0.1%	53.6%

Note: a_2 = direct effect of $X \rightarrow M_U$; b_2 = direct effect of $M_U \rightarrow Y$;
k = direct effect of $M_U \rightarrow M_O$; a_1 = direct effect of $X \rightarrow M_O$;
b_1 = direct effect of $M_O \rightarrow Y$; c = direct effect of $X \rightarrow Y$;
$\tilde{a}_1$ = direct effect of $X \rightarrow M_O$ with the omission of M_U;
$\tilde{b}_1$ = direct effect of $M_O \rightarrow Y$ with the omission of M_U;
$\tilde{c}$ = direct effect of $X \rightarrow Y$ with the omission of M_U.

[a] Hypothetical path coefficients for the model depicted in Figure 3.4.

Additionally, most cases in Table 3.1 generate positively biased values for the direct effect from X to Y, with only two exceptions of underestimation but the magnitude of bias is very small in both cases (-0.0002 and -0.0001). Additionally, it is very important to note Equation 3.3 also implies the bias in estimating c does not depend on the true value of c. From a practical perspective, this suggests even with a very large sample, omitting M_U can yield evidence of a medium positive (or a small negative) direct effect from X to Y when the actual direct effect is zero.

Table 3.2: Bias in the Estimated Direct Effect of X on Y and Estimated Indirect Effect via $M_O (k = 0)$.

Parameters[a]			Bias				Bias/True value			
a_2	b_2	k	$(\tilde{a}_1$ $-a_1)$	$(\tilde{b}_1$ $-b_1)$	$(\tilde{c}$ $-c)$	$(\tilde{a}_1\tilde{b}_1$ $-a_1 b_1)$	$(\tilde{a}_1$ $-a_1)$ $/a_1$	$(\tilde{b}_1$ $-b_1)$ $/b_1$	$(\tilde{c}$ $-c)$ $/c$	$(\tilde{a}_1\tilde{b}_1$ $-a_1 b_1)$ $/a_1 b_1$
0.1	0.1	0	0.00	0.00	0.0100	0.00	0.0%	0.0%	10.0%	0.0%
0.1	0.25	0	0.00	0.00	0.0250	0.00	0.0%	0.0%	25.0%	0.0%
0.1	0.4	0	0.00	0.00	0.0400	0.00	0.0%	0.0%	40.0%	0.0%
0.25	0.1	0	0.00	0.00	0.0250	0.00	0.0%	0.0%	25.0%	0.0%
0.25	0.25	0	0.00	0.00	0.0625	0.00	0.0%	0.0%	62.5%	0.0%
0.25	0.4	0	0.00	0.00	0.1000	0.00	0.0%	0.0%	100.0%	0.0%
0.4	0.1	0	0.00	0.00	0.0400	0.00	0.0%	0.0%	40.0%	0.0%
0.4	0.25	0	0.00	0.00	0.1000	0.00	0.0%	0.0%	100.0%	0.0%
0.4	0.4	0	0.00	0.00	0.1600	0.00	0.0%	0.0%	160.0%	0.0%

Note: a_2 = direct effect of $X \rightarrow M_U$; b_2 = direct effect of $M_U \rightarrow Y$;
k = direct effect of $M_U \rightarrow M_O$; a_1 = direct effect of $X \rightarrow M_O$;
b_1 = direct effect of $M_O \rightarrow Y$; c = direct effect of $X \rightarrow Y$;
$\tilde{a}_1$ = direct effect of $X \rightarrow M_O$ with the omission of M_U;
$\tilde{b}_1$ = direct effect of $M_O \rightarrow Y$ with the omission of M_U;
$\tilde{c}$ = direct effect of $X \rightarrow Y$ with the omission of M_U.
[a] Hypothetical path coefficients for the model depicted in Figure 3.4.

Tables 3.2 and 3.3 depict situations where the true underlying mediation process are parallel (Figure 3.1(e), $k = 0$) and serial (Figure 3.1(f), $a_1 = b_2 = 0$) two-mediator models, respectively. Importantly, all cases in Table 3.2 (parallel mediation model) yield positively biased direct effect from X to Y but unbiased indirect effect via M_O; while all cases in Table 3.3 (serial mediation model) yield positively biased indirect effect via M_O but unbiased direct effect from X to Y. These patterns are also implied by Equation 3.4 and 3.3, where $k = 0$ gives unbiased effect for $a_1 b_1$ in the parallel model and $b_2 = 0$ gives unbiased effect for c in the serial model. Additionally, the positive bias in both Table 3.2 and Table 3.3 increases as M_U-related path coefficients get larger. Most importantly, the actual indirect effect in the serial model is 0 (since $a_1 = 0$) but the omission of M_U can introduce bias to estimate a considerable indirect effect even when the researcher has a very large sample, especially when a_2 and k have large values in the true serial mediation model.

Table 3.3: Bias in the Estimated Direct Effect of X on Y and Estimated Indirect Effect via $M_O(a_1 = b_2 = 0)$.

Parameters[a]			Bias				Bias/True value			
a_2	b_2	k	$(\tilde{a}_1 - a_1)$	$(\tilde{b}_1 - b_1)$	$(\tilde{c} - c)$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1)$	$(\tilde{a}_1 - a_1) / a_1$	$(\tilde{b}_1 - b_1) / b_1$	$(\tilde{c} - c) / c$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1) / a_1 b_1$
0.1	0	0.1	0.01	0	0	0.0015	NA	0%	0%	NA
0.1	0	0.25	0.03	0	0	0.0038	NA	0%	0%	NA
0.1	0	0.4	0.04	0	0	0.0060	NA	0%	0%	NA
0.25	0	0.1	0.03	0	0	0.0038	NA	0%	0%	NA
0.25	0	0.25	0.06	0	0	0.0094	NA	0%	0%	NA
0.25	0	0.4	0.10	0	0	0.0150	NA	0%	0%	NA
0.4	0	0.1	0.04	0	0	0.0060	NA	0%	0%	NA
0.4	0	0.25	0.10	0	0	0.0150	NA	0%	0%	NA
0.4	0	0.4	0.16	0	0	0.0240	NA	0%	0%	NA

Note: a_2 = direct effect of $X \longrightarrow M_U$; b_2 = direct effect of $M_U \longrightarrow Y$;
k = direct effect of $M_U \longrightarrow M_O$; a_1 = direct effect of $X \longrightarrow M_O$;
b_1 = direct effect of $M_O \longrightarrow Y$; c = direct effect of $X \longrightarrow Y$;
$\tilde{a}_1$ = direct effect of $X \longrightarrow M_O$ with the omission of M_U;
$\tilde{b}_1$ = direct effect of $M_O \longrightarrow Y$ with the omission of M_U;
$\tilde{c}$ = direct effect of $X \longrightarrow Y$ with the omission of M_U.
[a] Hypothetical path coefficients for the model depicted in Figure 3.4.

Table 3.4 depicts situations where the omitted M_U is only a confounder for the mediator-to-outcome relation but not a mediator (i.e., $a_2 = 0$). In our illustrative example, this means the path coefficient from COND to IMPORT is zero. All cases in this table negatively bias the direct effect from X to Y and positively bias the indirect effect via M_O, which is consistent with previous literature (Fritz et al., 2016). By setting $a_2 = 0$ in Equation 3.1 through 3.3, we get the formulas for bias in estimating b_1 and c, which are the same as that in previous literature[9]. To further understand

[9] By setting $a_2 = 0$ in Equations 3.1 through 3.3, we can get that the bias in $\tilde{a}_1$ is 0, the bias in

Table 3.4: Bias in the Estimated Direct Effect of X on Y and Estimated Indirect Effect via $M_O (a_2 = 0)$.

Parameters[a]			Bias				Bias/True value			
a_2	b_2	k	$(\tilde{a}_1 - a_1)$	$(\tilde{b}_1 - b_1)$	$(\tilde{c} - c)$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1)$	$(\tilde{a}_1 - a_1) / a_1$	$(\tilde{b}_1 - b_1) / b_1$	$(\tilde{c} - c) / c$	$(\tilde{a}_1\tilde{b}_1 - a_1 b_1) / a_1 b_1$
0	0.1	0.1	0	0.01	-0.0021	0.002	0%	6.9%	-2.1%	6.9%
0	0.1	0.25	0	0.03	-0.0052	0.01	0%	17.4%	-5.2%	17.4%
0	0.1	0.4	0	0.04	-0.0083	0.01	0%	27.8%	-8.3%	27.8%
0	0.25	0.1	0	0.03	-0.0052	0.01	0%	17.4%	-5.2%	17.4%
0	0.25	0.25	0	0.07	-0.0130	0.01	0%	43.4%	-13.0%	43.4%
0	0.25	0.4	0	0.10	-0.0208	0.02	0%	69.4%	-20.8%	69.4%
0	0.4	0.1	0	0.04	-0.0083	0.01	0%	27.8%	-8.3%	27.8%
0	0.4	0.25	0	0.10	-0.0208	0.02	0%	69.4%	-20.8%	69.4%
0	0.4	0.4	0	0.17	-0.0333	0.03	0%	111.1%	-33.3%	111.1%

Note: a_2 = direct effect of $X \longrightarrow M_U$; b_2 = direct effect of $M_U \longrightarrow Y$;
k = direct effect of $M_U \longrightarrow M_O$; a_1 = direct effect of $X \longrightarrow M_O$;
b_1 = direct effect of $M_O \longrightarrow Y$; c = direct effect of $X \longrightarrow Y$;
$\tilde{a}_1$ = direct effect of $X \longrightarrow M_O$ with the omission of M_U;
$\tilde{b}_1$ = direct effect of $M_O \longrightarrow Y$ with the omission of M_U;
$\tilde{c}$ = direct effect of $X \longrightarrow Y$ with the omission of M_U.
[a] Hypothetical path coefficients for the model depicted in Figure 3.4.

this Table 3.4, we consider cases in this table as special situations of Table 3.1, where a_2 takes on the value of 0 in all cases. Implied by Equations 3.1, $a_2 = 0$ suggests an unbiased direct effect estimate from X to M_O and thus only the bias of the estimated direct effect from M_O to Y remains and contributes to the positively biased indirect effect via M_O. Implied by Equation 3.3, $a_2 = 0$ indicates a negative bias for the estimated direct effect from X to Y, under the conditions that a_1, k

$\tilde{b}_1$ is $b_2 \cdot k/(1 - a_2{}^2)$, and the bias for the bias in $\tilde{c}$ is $b_2 \cdot (-k \cdot a_1)/(1 - a_2{}^2)$. These are exactly the same as the bias formula presented in Fritz et al. (2016) (p. 684), where they wrote b_2 as e_1 and they use C_1 to denote the confounder. Their $r_{C_1 M}$ is our $\rho_{M_O M_U}$, and their r_{XM} is our ρ_{XM_O}. The Appendix presents the formulas of correlations as a function of path coefficients. When $a_2 = 0$, we get $\rho_{M_O M_U} = k$ and $\rho_{XM_O} = a_1$. Thus, their bias in estimating b_1, which is $e_1 \cdot \left[\frac{r_{C_1 M}}{1 - r_{XM}^2} \left(\frac{s_{C_1}}{s_X}\right)\right]$, is equivalent to our $b_2 \cdot \frac{k}{(1 - a_2{}^2)}$; their bias in estimating c, which is $e_1 \cdot \left[\frac{-r_{XM} r_{C_1 M}}{1 - r_{XM}^2} \left(\frac{s_{C_1}}{s_X}\right)\right]$, is equivalent to our $b_2 \cdot \frac{-k \cdot a_1}{1 - a_2{}^2}$.

and b_2 are all positive.

3.8 Correlation Framework and Sensitivity Analysis for Omitted Alternative Mediators

We have demonstrated that omitting alternative mediators/post-treatment confounders may bias the estimate of the direct effect from X to Y as well as the indirect effect via M_O. The direction and magnitude of bias can vary substantially under different underlying mediating processes. From a practical perspective, we never know the actual parameters in the true model, but we can use our analysis to quantify the robustness of any inference regarding M_O to a potential post-treatment confounder.

Sensitivity analyses can serve as a useful tool to inform the strength of evidence for specific inferences by quantifying the conditions that would alter the inference (e.g., Frank, 2000; Imbens, 2003; Rosenbaum, 2002; VanderWeele & Arah, 2011). For example, if a study focuses on estimating a treatment effect, sensitivity analyses can generate statements about how strong an omitted variable would have to be correlated with the treatment and with the outcome to invalidate an inference of an effect of the treatment on the outcome. As such, recent approaches to sensitivity analysis help interpreters of research quantify the conditions necessary to invalidate an inference drawing on familiar quantities such as correlations (Frank, 2000), percentage of variance explained (Cinelli & Hazlett, 2018) or graphical representations such as contour plots (Imbens, 2003).

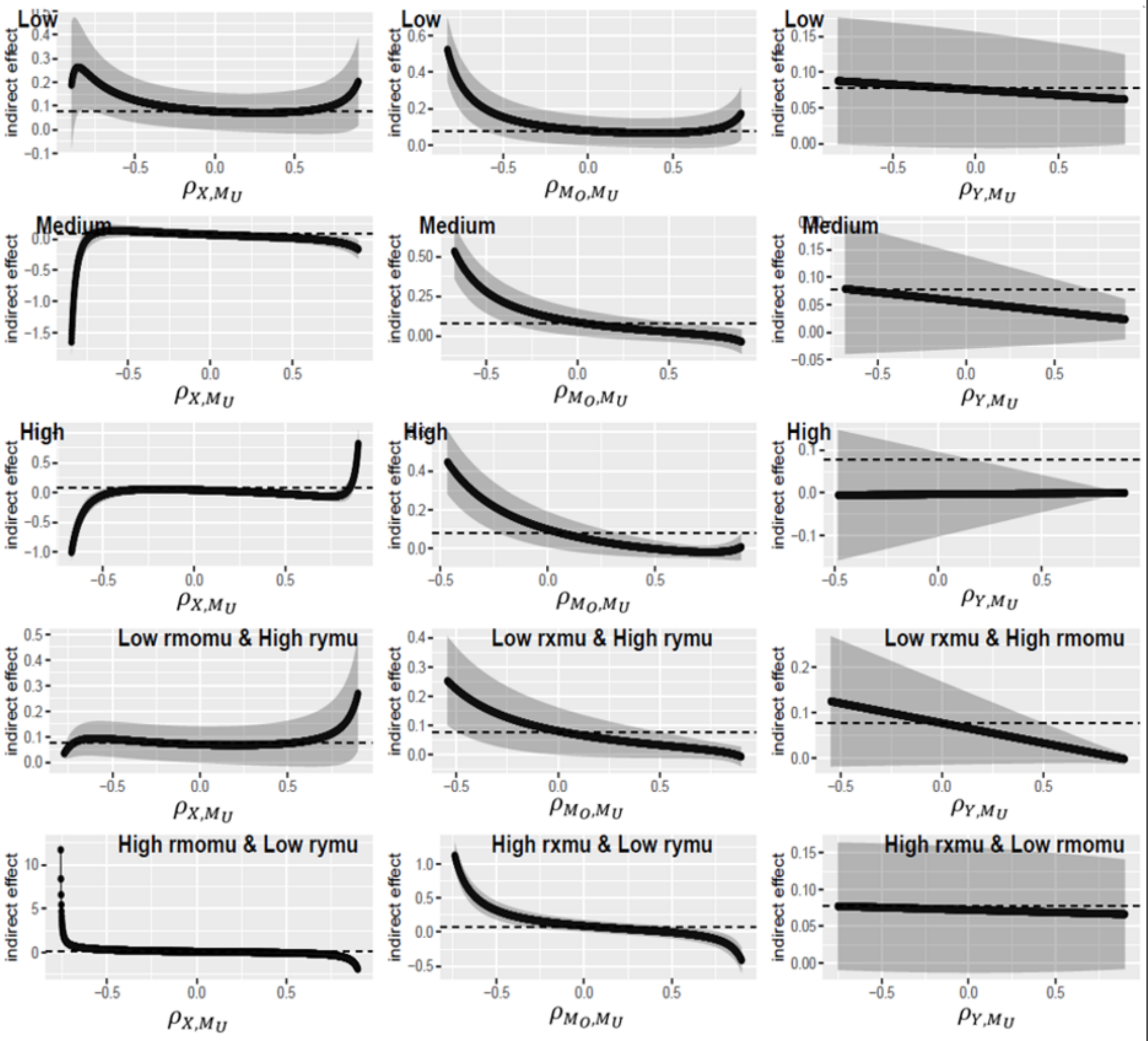
In this section, we quantify the robustness of inferences by evaluating how sensitive the estimated direct and indirect effects of M_O are to a potential post-treatment confounder M_U. For example, we can quantify how large must be the effect of the M_U on the M_O and Y to change a statistically significant direct or indirect effect to zero. Specifically, we quantify the sensitivity of estimates and robustness of inferences as functions of correlations between M_U and other observed variables X, Y and M_O, so that empirical researchers can use these quantities to discuss the robustness of their inferences in terms of M_U.

To introduce this sensitivity analysis approach, we first present the correlation framework: based on Equations 3.1 through 3.4 we write the bias in estimating a_1, b_1, and c as function of correlations among the four variables X, M_O, M_U and Y: ρ_{X,M_O}, ρ_{X,M_U}, $\rho_{X,Y}$, ρ_{Y,M_O}, ρ_{Y,M_U} and ρ_{M_O,M_U}. The Appendix provides details and the final derivation results. Compared to the parameter framework Equation 3.1 through Equation 3.4 where all the path parameters are unknown (i.e., a_1, b_1, c, a_2, b_2 and k) to substantive researchers, now we can always estimate ρ_{X,M_O}, $\rho_{X,Y}$, and ρ_{Y,M_O} from the sample data. As such, the bias only depends on the three unknown M_U-related correlations: ρ_{X,M_U}, ρ_{Y,M_U} and ρ_{M_O,M_U}. These three unknown correlations are the key parameters in our sensitivity analysis approach.

To demonstrate this sensitivity analysis approach, we go back to our illustrative example regarding presumed influence of media use. Assume we fit the model only with one mediator PMI, where the results were presented in the right panel of Figure 3.2. The specific indirect effect via PMI was estimated as 0.078 and significantly positive. Next, we ask how sensitive this estimate is to an unobserved post-treatment confounder/mediator. For example, we consider IMPORT as a potential confounder, but we fail to measure this variable. Focusing on the indirect effect via PMI, Figure 3.8 presents the result for the sensitivity analyses based on the three correlations as sensitivity parameters: the correlations between the unobserved post-treatment confounder M_U and the other observed variables X, Y and M_O.

Considering that we have three sensitivity parameters, we include 15 scenarios in Figure 3.8 to comprehensively present the sensitivity analyses. In all scenarios, the horizontal dotted line is drawn at the point estimate of 0.078, which is the estimated indirect effect via PMI when we fit the model excluding the confounder M_U. The solid black line plots the estimated indirect effect via PMI against differing values of one correlation related to the confounder M_U. The grey region represents the 95% confidence interval based on the Delta method (Sobel, 1982) [10]. Each column includes five scenarios plotting how the estimated indirect effect varies with one particular M_U-related correlation: the left column has x-axis representing the correlation between X and M_U; the middle column has x-axis representing the correlation between M_O and M_U; the right column has x-axis representing the correlation between Y and M_U. Each row includes three scenarios where

[10] The standard error of the indirect effect is approximated by $\sqrt{a_1{}^2 var\,(b_1) + b_1{}^2 var\,(a_1)}$.



Note: The dashed line represents the estimated indirect effect for no omitted post-treatment confounder. The grey area represents the 95% confidence interval for the indirect effect via observed mediator at each value of the unobserved M_U-related correlation. The solid line represents the estimated indirect effect via observed mediator at each value of the unobserved M_U-related correlation.

Figure 3.8: Sensitivity analysis for unobserved post-treatment confounder.

the other two unmanipulated M_U-related correlations are fixed at a certain level: the 1^{st} to 3^{rd} rows show scenarios where the two correlations taking on low, medium and high values[11] , respectively; the 4^{th} and 5^{th} rows show scenarios where the two correlations taking on one low, one high and one high, one low values, respectively.

We interpret the sensitivity analyses by looking at each column separately. The five plots in the left column show how the estimated indirect effect changes with different values of the correlation between COND and the potential posttreatment confounder, when the correlation between PMI and the confounder M_U (rmomu) and the correlation between REACTION and the confounder M_U (rymu) are fixed at different values. For example, the fourth plot in the left column indicates when the correlation between PMI and M_U (rmomu) is low (0.1) and the correlation between REACTION and M_U (rymu) is high (0.5), the original inference about the direction of the indirect effect via PMI would always be maintained, no matter how large the correlation between COND and M_U. However, if we consider sampling variability, the confidence interval covers the value of zero when the correlation between COND and M_U is between -0.043 and 0.834. This implies that the conclusion of a positive indirect effect via PMI is sensitive to a posttreatment confounder. Similar conclusions can be drawn based on other plots in the first column: although only extreme values of the correlation between COND and M_U can alter the direction of the indirect effect from positive to negative, once we take into account the sampling variability, the confidence interval can cover zero even when the correlation between COND and M_U is around 0. The five plots in the middle column imply the same conclusion: the direction of the indirect effect via PMI can be maintained in most cases unless the correlation between PMI and M_U takes on some extreme values; however, once sampling variability is taken into account, the confidence interval can cover zero even when the correlation between PMI and M_U is close to zero. The scenarios in the right column, which plot the estimated indirect effect against differing values of the correlation between

[11] We used 0.1, 0.3 and 0.5 for low, medium and high correlations between continuous variables (i.e., M_O, M_U and Y). For correlations involving the binary variable X, we used 0.2, 0.5 and 0.8 as low, medium and high Cohen's D, which corresponds to correlations of 0.0995, 0.2425 and 0.3713 as low, medium and high, respectively.

REACTION and M_U, support an even stronger conclusion that the positive indirect effect via PMI is very sensitive to a post-treatment confounder M_U: in all five plots, the confidence interval can cover zero for most values of the correlation between REACTION and M_U, especially when the other two M_U-related correlations are high (the 3^{rd} plot).

3.9 Discussion

The major conclusion of this article is that omitting a mediator M_U will typically generate biased (more precisely, inconsistent) estimates for the specific indirect effect via M_O, as well as the direct effect from X to Y. Additionally, the magnitude of bias (inconsistency) can be substantial. In our illustrative example that studies whether media (X) affects people's attitudes or behaviors (Y) through changing people's perceptions regarding how other people may be influenced by the media (M_O), excluding the alternative mediator of the perceived importance of the article (M_U) has given us a different conclusion that the indirect effect via presumed media influence (M_O) is significantly positive. Once the perceived media importance (M_U) was included, the indirect effect via presumed media influence (M_O) is not significant from 0, decreasing from 0.078 to 0.045. Though the inference about the direct path from the direct path from article location (X) to participants' reaction (Y) did not change no matter whether the perceived media importance (M_U) was included or not, the point estimate decreased from 0.082 to 0.033 once we included the M_U in our model.

The exact pattern of bias (inconsistency) depends on the specific underlying mediation process. In the parallel two-mediator model (Figure 3.1(e)) where the omitted mediator is independent of the observed mediator, the estimate of the specific indirect effect via M_O is not biased but the estimate of the direct effect from X to Y can be either positive or negatively biased, depending on the specific indirect effect via the omitted mediator M_U. If the specific indirect effect via omitted mediator M_U is positive, then the direct path from X to Y is overestimated. Some credits via the direct effect from X to Y should be attributed to M_U as a mediator. If the specific indirect effect via omitted mediator M_U is negative, then the direct path from X to Y is underestimated. The magnitude of the bias (inconsistency) in estimating c is exactly the true specific indirect effect via M_U.

In the serial two-mediator model (Figure 3.1(f)), the estimate of the direct effect from X to Y is unbiased but the estimate of the specific indirect effect via M_O can be either positively or negatively biased, depending on the path coefficients directly related to the unobserved mediator M_U. The estimated effect attributed to $X \rightarrow M_O$ should be attributed to $X \rightarrow M_U \rightarrow M_O$. Note the true path from X to M_O is zero, indicating that the true specific indirect effect via M_O is zero. By excluding M_U, we may conclude a nonzero indirect effect via M_O.

If the underlying mediating process is a more general two-mediator model where all path coefficients are positive, then the indirect effect via M_O is overestimated and the magnitude of bias can vary substantially. The estimate of the direct effect from X to Y can be either positively or negatively biased. When the path from M_U to M_O (k) is relatively small, we overestimate the direct effect (c) while when k gets close to 1, we underestimated the direct effect (c). Further, the larger the path coefficient from M_U to M_O (k), the larger the positive bias in estimating the indirect effect via M_O. Importantly, the conditions to obtain unbiased direct and indirect effect estimates are not the same, so that one cannot simultaneously obtain accurate direct and indirect effects. Situations will become even more complicated once we consider sampling variability. Omitting an alternative mediator can have either have no implications or disastrous consequences for hypothesis testing regarding the observed mediator M_O. Therefore, we propose a sensitivity analysis approach where the three M_U-related correlations serve as sensitivity parameters. We are also developing an easy-to-use R package for empirical researchers to examine the robustness of their inference regarding M_O to a potential omitted mediator M_U. The demonstration for the sensitivity analysis in this paper only shows one way this package can be applied. If the researcher has any knowledge about a certain M_U, they can choose to specify two of the unknown M_U-related correlations so that they can focus on one figure to see how the other M_U-related correlation affects the estimated effect and the inference.

One may ask that what if more than one mediator is omitted? From a practical perspective, we can never know the true underlying mediating process and there can always be another omitted mediator. We argue that for the purpose of quantifying the strength of evidence in making inference, it is enough to consider one mediator that captures all sources of bias. To better conceptualize the omitted mediator in our thought experiment of the sensitivity analysis, we can think about M_U as one latent variable that captures all potential omitted mediators that may bias our inference for M_O.

Finally, it is important to note that the definition of the indirect effect and direct effect is different from the natural indirect effect (NIE), natural direct and the control direct effect in the counterfactual framework. As reviewed before, this distinction can be considerable under models with multiple mediators, which emphasized the importance to clarify the definition of indirect effect and direct effect in applied research. Additionally, although the counterfactual framework allows us to relax those parametric assumptions, we argue that, by specifying a parametric model, we can see how omitting another mediator may bias our estimation for each specific path coefficient we are interested in (i.e., a_1, b_1 and c).

3.10 Limitations and Future Directions

Several limitations of the current work suggest avenues for future research. First, we have strong assumptions for model specifications, including no mediator-outcome interaction and the original conditions (X) are randomly assigned. A more complex situation arises when these assumptions are relaxed. Second, we applied the delta method to approximate the sampling variability, which may be not accurate under some scenarios. Third, we acknowledge that three sensitivity parameters are a lot to consider. Though we provided several plots in the sensitivity approach to accommodate different scenarios, it would be valuable if future studies can reduce the number of sensitivity parameters. Furthermore, we only examined the cross-sectional dual-mediator model. Recent research has suggested longitudinal designs to test mediation because cross-sectional examination of mediation can generate biased estimates (e.g., Maxwell & Cole, 2007; Maxwell, Cole, & Mitchell, 2011; Mitchell & Maxwell, 2013). Accordingly, future studies should consider whether omitting another mediator may bias our estimation in longitudinal designs.

CHAPTER 4

APPLYING A PARAMETER FRAMEWORK TO QUANTIFY INCONSISTENCY IN A TIME VARYING MEDIATION MODEL

4.1 Introduction

This chapter will continue the discussion in Chapter 3 about an unobserved mediator as a posttreatment confounder in a single-mediator model. Specifically, I will leverage the parameter framework to further discuss how inconsistency is generated for each path coefficient we are interested in when a posttreatment confounder is omitted. Therefore in the first section of this chapter I will develop a parameter framework for characterizing inconsistency in mediation models. Then in the second part I will apply this parameter framework to a longitudinal design.

Two important tools will serve as the basis for all the discussion in this chapter: namely the Law of Iterated Expectation (LIE) and the Linear Regression framework. These powerful tools allow us to grasp a deep and intuitive understanding about the mechanisms underlying inconsistency generation. More importantly, this understanding can be applied in several ways: (1) it helps us better understand the patterns of how each M_U-related parameter affects inconsistency; (2) it has implications for how we may consider the effects of the unobserved confounder M_U on the indirect effect via M_O and the direct effect from X to Y; (3) this understanding also goes beyond the cross-sectional model and depicts the underlying story of a post-treatment confounder in a longitudinal single-mediator model as well.

4.2 Deriving Inconsistency Using the Law of Iterated Expectation for a Parameter Framework

The Law of Iterated Expectation (LIE) states that $E(Y|X) = E[E\left(Y|X,Z\right)|X]$ where X, Y and are Z three variables and $E\left(Y|X\right)$ represents the conditional expectation (or conditional mean) of Y given X (Wooldridge, 2009). The LIE allows us to apply a two-step approach to solve the conditional mean: $E(Y|X)$. First, we find $E\left(Y|X,Z\right)$, which is the conditional mean of Y given

X and Z. As such, we get a function of X and Z for this conditional mean. Second, we use the information from $E(Z|X)$, which is a function of X, to solve the expected value of $E(Y|X,Z)$ conditional on X.

The LIE can be a very useful tool to derive inconsistency at the population level when we omit a related independent variable in a regression model (model misspecification). For example, consider that the true model is $E(Y|X,Z) = aX + bZ$ and the correlation between X and Z is not zero. But when we fit a regression model with all population data, we omit Z and only regress Y on X. This can lead to an inconsistent estimator for the parameter a. In this scenario, we can apply the LIE to derive the inconsistency. We do this by writing $E(Y|X) = E[E(Y|X,Z)|X] = a \cdot X + b \cdot E(Z|X)$ and then plugging in $E(Z|X)$ (which is a function of X) to solve the question. The underlying intuition is that part of the explanatory credit that belongs to the omitted Z has been allocated to X. As a result, the mistakenly attributed credit generates inconsistency when we estimate a while omitting Z.

To note, the derivation here is at the population level and therefore is free of sampling error. Technically speaking, the difference between $\tilde{a}_1$, $\tilde{b}_1$,$\tilde{c}$ and a_1, b_1, c are *inconsistencies*, not *bias*. That is, $\tilde{a}_1$, $\tilde{b}_1$, $\tilde{c}$ are the estimates we get even when we have all population data available (i.e., sample size $n \to \infty$). In all following discussion, “bias” in the population parameters is a rough way to describe “inconsistency” for an audience more comfortable with conceptualizations of bias.

Recall that the true model and the model that omits the unobserved mediator can be written as follows, in Figure 3.4 (from Chapter 3). X is the exposure variable, M_O is the observed mediator of interest, Y is the outcome, and M_U is the unobserved mediator. We are interested in estimating the specific indirect effect via M_O and the direct path from X to Y. As such, three true effects are of key interest: a_1 that represents the path $X \to M_O$, b_1 that represents the path $M_O \to Y$, and c that represents the path $X \to Y$. When M_U is omitted, $\tilde{a}_1$, $\tilde{b}_1$ and $\tilde{c}$ are the estimated effect for $X \to M_O$, $M_O \to Y$ and $X \to Y$, respectively.

Now we apply LIE to express the inconsistency due to omitting M_U. That is, we want to derive the differences between the estimators and their true effects when the confounder M_U is present yet not included, and we have the population data (i.e., sample size $n \to \infty$). In this case, the omitted M_U plays the role of Z in our previous example. Specifically, we first write our true models as Equations 4.1 through 4.3:

$$E(M_O|M_U, X) = k \cdot M_U + a_1 \cdot X \tag{4.1}$$

$$E(M_U|X) = a_2 \cdot X \tag{4.2}$$

$$E(Y|M_O, M_U, X) = b_1 \cdot M_O + b_2 \cdot M_U + c \cdot X \tag{4.3}$$

By LIE, we can further write equations 4.4 to 4.5 to see what happens when M_U is excluded:

$$E(M_O|X) = E\left[E\left(M_O|M_U, X\right)|X\right] = k \cdot E(M_U|X) + a_1 \cdot X \tag{4.4}$$

$$E(Y|M_O, X) = E\left[E\left(Y|M_O, M_U, X\right)|M_O, X\right] = b_1 \cdot M_O + b_2 \cdot E(M_U|X, M_O) + c \cdot X \tag{4.5}$$

Now write:

$$E\left(M_U|X\right) = \beta_1 \cdot X \tag{4.6}$$

$$E\left(M_U|X,\ M_O\right) = \beta_2 \cdot M_O + \beta_3 \cdot X \tag{4.7}$$

Plugging Equation 4.6 and 4.7 back to Equation 4.4 and 4.5, we can get formulas for inconsistencies in $\tilde{a}_1$, $\tilde{b}_1$ and $\tilde{c}$, as follows:

$$\tilde{a}_1 - a_1 = k \cdot \beta_1 \tag{4.8}$$

$$\tilde{b}_1 - b_1 = b_2 \cdot \beta_2 \tag{4.9}$$

$$\tilde{c} - c = b_2 \cdot \beta_3 \tag{4.10}$$

Importantly, as implied by Equation 4.6 and 4.7, β_1, β_2 and β_3 are three regression coefficients (again, at the population level). Specifically, β_1 is the regression coefficient of X when we regress M_U on X, which is also equivalent to a_2. β_2 and β_3 are regression coefficients of M_O and X, respectively, when regressing M_U on M_O and X.

Before going to deep discussion about the inconsistency, it is necessary to clarify that I use the linear regression framework here in a “loose” way to serve as a tool for easy interpretation only. No causality is implied in the regression models. For example, Equation 4.7 implies a linear regression model where M_O and X are predictors for M_U, which only has statistical meaning but no causal implications. In other words, the coefficients β_2 and β_3 only represent statistical association between M_O, X and M_U. They do not imply that M_O and X cause M_U.

4.3 Understanding What Happens When Omitting M_U

Now I will present how Equations 4.8 to 4.10 can help us understand the underlying mechanisms of sources of bias. To simplify the discussion, we assume all parameters are positive. Chapter 3 shows that the bias in estimating a_1 and b_1 are always positive. But the direction of $\tilde{c}$'s bias depends on the relative magnitude of ρ_{X,M_U} and $\rho_{X,M_O} \cdot \rho_{M_O,M_U}$. As we will see later, this is much more than just a mathematical result. In fact, the discussion for $\tilde{c}$'s story has inspired me to consider more deeply about what's actually happening when omitting M_U. Importantly, this parameter framework allows us to have a powerful tool to understand how the omittance of M_U generates the bias. Following Figure 4.1 presents a summary of this interpretative framework, in which the formulas in the right columns are based on Equations 4.8 through 4.10.

We will start our consideration of this framework from the first row for $\tilde{a}_1$, which happens to be the most direct and easy result. The formula tells us that the bias is the product of k and a_2. From the causal pathways in the left column, we can see that this bias comes from the causal pathway $X \to M_U \to M_O$. That is, $\tilde{a}_1$ measures the effect $X \to M_O$ plus the effect $X \to M_U \to M_O$. Therefore, omitting M_U is basically omitting a mediator when estimating a_1.

The stories for $\tilde{b}_1$ and $\tilde{c}$ are more complicated. First we observe that the bias in estimating b_1 and c are both a weighted version of b_2, which represents the causal pathway $M_U \to Y$. Interestingly enough, the two weights for the bias in $\tilde{b}_1$ and $\tilde{c}$ are the two partial regression coefficients from one regression model. The last row of Figure 4.1 presents this crucial regression model, in which we use X and M_O to predict M_U. Specifically, the weight for $\tilde{b}_1$'s bias is the coefficient for M_O and the weight for $\tilde{c}$'s bias is the coefficient for X. We know that a partial regression coefficient in a multiple regression model measures the unique contribution of the predictor to predict/explain the variance

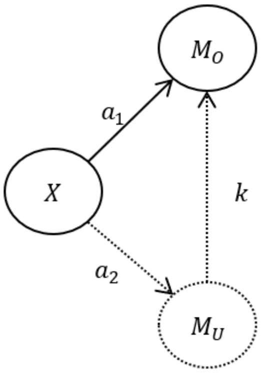
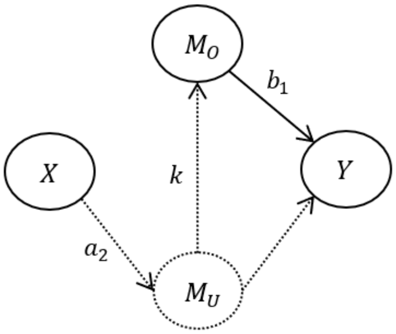
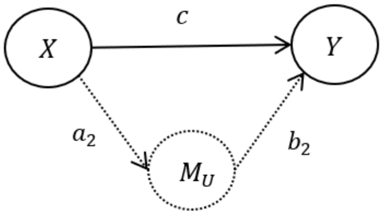
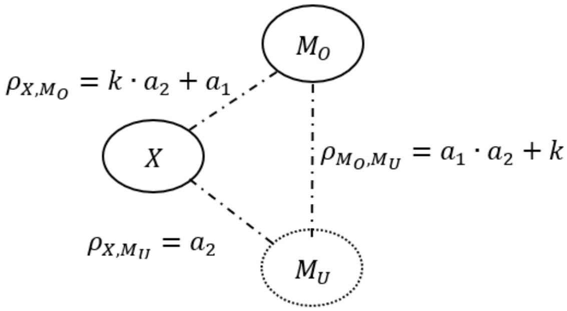
 Omitting the dashed paths $X \rightarrow M_U \rightarrow M_O$leads to bias for the solid path $X \rightarrow M_O$.	$\tilde{a}_1 = a_1 + k \cdot a_{2,}$ • k plays a role in $k \cdot a_2$; • a_2 plays a role in $k \cdot a_2$; • no role of b_2.
 Omitting the dashed paths $X \rightarrow M_U \rightarrow M_O$leads to bias for the solid path $M_O \rightarrow Y$.	$\tilde{b}_1 = b_1 + b_2 \cdot \beta_{2,}$ where β_2 is the coefficient of M_O in the key regression model; • k affects by positively influencing β_2; • a_2 affects through β_2; • b_2 plays a role in $b_2 \cdot \beta_2$.
 Omitting the dashed paths $X \rightarrow M_U \rightarrow Y$ leads to bias for the solid path $X \rightarrow Y$.	$\tilde{c} = c + b_2 \cdot \beta_{3,}$ where β_3 is the coefficient of X in the key regression model; • k affects by negatively influencing β_3; • a_2 affects through β_3; • the role of b_2 depends on the direction of β_3.
The key regression model: $E(M_U \mid M_O, X) = \beta_2 \cdot M_O + \beta_3 \cdot X$	 M_O $\rho_{X,M_O} = k \cdot a_2 + a_1$ X $\rho_{M_O,M_U} = a_1 \cdot a_2 + k$ $\rho_{X,M_U} = a_2$ M_U

Figure 4.1: Parameter framework to understand what happens when omitting M_U.

of the outcome (again, no causality is implied here). In this context, the unique contribution made by M_O to explain M_U composes the weight applied to b_2 to generate the bias in estimating b_1. On the other hand, the unique contribution attributed by X to explain M_U becomes the weight of b_2 to form the bias c.

This observation that centers on regressing M_U on X and M_O actually points us to understand two competing mechanisms about how omitting M_U affects the estimation of b_1 and c. If we trace all the causal pathways that go through M_U, the beginning point is surely $X \rightarrow M_U$. But then this causal pathway splits into two parts. The first part goes through M_O to the ending point of Y. This first part passes through the b_1 pathway and thus it contributes to the bias in estimating b_1. The second part, on the other hand, passes directly to Y after M_U. The second part here $X \rightarrow M_U \rightarrow Y$ picks up a mediation pathway from X to Y and as a result, it is responsible for bias in estimating c.

Importantly, the role of the omitted M_U is different in these two parts though confounding and mediation are known to have statistical similarities (MacKinnon, Krull, & Lockwood, 2000). In the first part for the bias in $\tilde{b}_1$, M_U is a real confounder because it has an effect on both M_O and Y. In other words, the explanatory credit that belongs to M_U via direct effects $M_U \rightarrow M_O$ and $M_U \rightarrow Y$ are allocated to the direct effect $M_O \rightarrow Y$. This explains why $\tilde{b}_1$ is the sum of b_1 and $b_2 \cdot \beta_2$. Intuitively, we can consider β_2 as a measure for effects that flow through $X \rightarrow M_U \rightarrow M_O$. On the contrary, M_U is only a mediator instead of confounder for estimating c. As the third row in Figure 4.1 shows, the estimated effect of c is biased because we count the indirect effects $X \rightarrow M_U \rightarrow Y$ as part of the direct effect $X \rightarrow Y$. Because part of the effect from $X \rightarrow M_U$ flows to M_O, we only get the remaining part multiplied by b_2 to form the bias in estimating c. The remaining part of $X \rightarrow M_U$ is exactly what is represented by β_3.

Because these two flows (i.e., $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$) both orginate from $X \rightarrow M_U$, they are in fact competing with the path $X \rightarrow M_U$. This competition manifests iteself by the linear regression model in our previous discussion, in which M_U is regressed on X and M_O. In other words, X and M_O are competing with each other to explain the variation in M_U. The part that X wins then continues to the flow toward Y (i.e., $X \rightarrow M_U \rightarrow Y$), which results in the bias in estimating c. On the other hand, the unique contribution made by M_O passes through M_O to Y (i.e., $X \rightarrow M_U \rightarrow M_O \rightarrow Y$), which leads to the bias in estimating b_1.

Equipped with this framwork, we are now able to have a more intuitive understanding for the direction of bias in $\tilde{c}$ discussed in Chapter 3. Previously, we said that the sign of the bias in estimating c depends on the relative magnitude of ρ_{X,M_U} versus $\rho_{X,M_O} \cdot \rho_{M_O,M_U}$. If $\rho_{X,M_U} > \rho_{X,M_O} \cdot \rho_{M_O,M_U}$, we overestimat c. Otherwise, we underestimate c when $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$. The comparison between ρ_{X,M_U} and $\rho_{X,M_O} \cdot \rho_{M_O,M_U}$ reflects exaclty the competition between X and M_O to explain the variation in M_U. This is because $\left(\rho_{X,M_U} - \rho_{X,M_O} \cdot \rho_{M_O,M_U}\right)$ is just the numerator of the regression coefficient for X. Alternatively, this is also equivalent to the partial correlation between X and M_U conditional on M_O. When $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$ (and assuming all correlations are positive), we have the unconditional (zero-order) correlation between X and M_U reversing the sign once conditional on M_O. This effect is also known as suppression in the literature (Cohen & Cohen, 1983) and somestimes M_O is called a distorter variable (Rosenberg, 1968). That is, the relationship between X and M_U gets suppressed or distorted once we include M_O to explain the outcome M_U. Intuitively, we may consider this as X losing the competion with M_U to explain M_O. When this is the case, we underestimate c. (Again, we only refer to statistical relations rather than any causal relationships when talking about a suppressor.)

Because parameters represent causal pathways, we see that the parameter framework allows us to tell stories about mechanisms generating the bias. In the following section, more analyses will be presented that applies this framework to better understand how M_U-related parameters affect the magnitude of bias. As we will see, this framework (Figure 4.1) serves as a very useful instrument in interpreting how bias vary under different conditions, providing us a deeper discussion compared to that in Chapter 3.

4.4 Using the Mechanism to Understand How Inconsistency Changes with M_U-related Parameters

4.4.1 How inconsistency changes with different levels of k.

We start with the causal effect from M_U to M_O, which is represented as k. In Chapter 3, we have mentioned that the magnitude of bias in estimating both a_1 and b_1 increase as the causal effect k becomes stronger. But now we have a framework to better understand why this is the case. Focusing on the effects $X \longrightarrow M_U \longrightarrow M_O$ as shown in the first row of Figure 4.1, larger k leads us to allocate more explanatory power to M_O that in fact belongs to M_U as a mediator. Therefore, we overestimate a_1 and the difference between $\tilde{a}_1$ and a_1 grows as k gets bigger. Similary, k plays a role in the chain of causal effects $X \longrightarrow M_U \longrightarrow M_O \longrightarrow Y$ (the second row of Figure 4.1). Then larger k leads more explanatory power via M_U to be attributed to M_O, which manifests as larger bias in estimating b_1. Thus, bias in both $\tilde{a}_1$ and $\tilde{b}_1$ increase, leading to more serious overestimation of the indirect effect $X \longrightarrow M_O \longrightarrow Y$.

Alternatively, we can interpret why the bias in $\tilde{b}_1$ increases with k by considering the role of k in the key regression model discussed before. In Figure 4.1 (last row), this key regression model is presented with the zero-order bivariate correlations as functions of parameters. We can tell k plays an important role in the correlation between M_U and M_O but the correlation between M_U and X only depends on a_2. In the example of Figure 3.5 (from Chapter 3), we fix the value of a_2 as 0.2 while increasing the value of k. That is, the correlation between M_U and M_O becomes larger and larger compared to that between M_U and X. This leads to greater importance attached to M_O in the competition between X and M_O to explain M_U, which also means that X becomes less and less important. Correspondingly, the partial regression coefficient for $M_O(\beta_2)$ becomes larger while the partial regression coefficient for $X(\beta_3)$ turns to be smaller as k increases. Based on previous discussion, we know that β_2 is responsible for the bias in $\tilde{b}_1$ and β_3 is responsible for the bias in $\tilde{c}$. Therefore, the increase in β_2 explains the growing bias in $\tilde{b}_1$ and the decrease in β_3 explains the declining bias in $\tilde{c}$.

Moreover, in Chapter 3 we also emphasized that after the positive bias in $\tilde{c}$ falls to 0 it continues decreasing to negative values as k increases to 1. That is, when k is relatively small we overestimate c while when k becomes larger we underestimate c. And as k gets closer to 1, the magnitude of negative bias in estimating c keeps growing. This is in fact consistent with our previous discussion regarding the direction of $\tilde{c}$'s bias. As k becomes larger, suppression starts to take place and becomes increasingly serious. That is, the difference between ρ_{X,M_U} and $\rho_{X,M_O} \cdot \rho_{M_O,M_U}$ becomes larger. Again, we can apply our competition story to interpret this observation: as k becomes larger, X becomes weaker and weaker in its competition with M_O. This leads more effect from $X \rightarrow M_U$ to be attracted towards $M_O \rightarrow Y$ rather than directly flowing to Y, which means that the overestimation of the direct effect $M_O \rightarrow Y(b_1)$ keeps growing. Meanwhile, as less effect remains to pass through $X \rightarrow M_U \rightarrow Y$, the bias in $\tilde{c}$ first decreases to 0 and then further becomes more and more negative.

4.4.2 How inconsistency changes with different levels of a_2

In Chapter 3, we discussed the larger the direct effect $X \rightarrow M_U(a_2)$, the larger the bias is estimating a_1. Relying on the first row in Figure 4.1, we are able to explain this as the indirect effect $X \rightarrow M_U \rightarrow M_O$ is allocated to $X \rightarrow M_O$, where the omitted M_U plays a role of mediator.

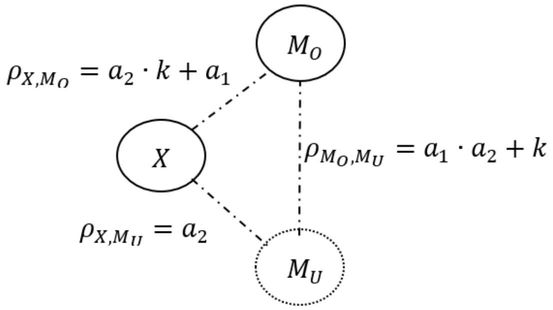
In Chapter 3, we also discussed that the effects of a_2 on the bias in $\tilde{b}_2$ and $\tilde{c}$ rely on the values of other parameters. Specifically, Figure 3.6(a) in Chapter 3 gives the scenario where the magnitude of a_1 and k are both relatively small (with values of 0.15 and 0.22) while in Figure 3.6(b) a_1 and k are both relatively large (with values of 0.6 and 0.5).

Figure 4.2 presents how the competing story in the parameter framework we discussed before for the bias in estimating b_1 and c can be served as a useful tool to interpret the distinctions between Figure 3.6(a) and 3.6(b). We know that the direct effect $X \rightarrow M_U$ (a_2) splits into two causal pathways after the point of M_U, one toward M_O and the other towards Y, generating the bias in $\tilde{b}_1$ and $\tilde{c}$, respectively. And the way the effect $X \rightarrow M_U$ (a_2) gets split can be understood by a competition between X and M_O to predict M_U. Interestingly, how the competition is affected by

a_2 depends on the magnitudes of a_1 and k, which is exactly the distinction between Figure 3.6(a) and 3.6(b).

The key regression model:

$$\boldsymbol{E(M_U|M_O,X) = \beta_2 \cdot M_O + \beta_3 \cdot X}$$


M_O
$\rho_{X,M_O} = a_2 \cdot k + a_1$
X
$\rho_{M_O,M_U} = a_1 \cdot a_2 + k$
$\rho_{X,M_U} = a_2$
M_U

a_2 starts from small value	as a_2 becomes larger	Affecting *b₁* through β_2	Affecting c through β_3	Comment
a_1 and k are small		M_O gains less importance; β_2 decreases	X gains more importance; β_3 increases	c can be first negatively biased then becomes positively biased (bias first reduce then increase)
a_1 and k are large		M_O gains more importance; β_2 increases	X gains less importance; β_3 decreases	c is always negatively biased (more seriously biased along the way)

Figure 4.2: Understand how bias changes with different levels of a_2.

In Figure 3.6(a), a_1 and k are relatively small, the increase in a_2 has larger effect on ρ_{X,M_U} compared to ρ_{X,M_O} and ρ_{M_O,M_U}. This is because the latter two correlations can be read as a_2 weighted by a_1 and k. That is, the increase in a_2 needs to be discounted when reflected in ρ_{X,M_O} and ρ_{M_O,M_U}. In comparison, ρ_{X,M_U} can get the 100% growth from a_2 because ρ_{X,M_U} is just equivalent to a_2. Therefore, when a_1 and k are small in Figure 3.6(a), the increase in a_2 leads X to gain more and more advantage in the competition with M_O. In this context, larger a_2 causes more explanatory credit to be allocated to $X \to M_U \to Y$ with less credit attributed to $X \to M_U \to M_O \to Y$. As such, though c can be underestimated in the very beginning (a_2 is very close to 0), the negative bias becomes positive quickly and keeps growing larger. At the same time, the overestimation in $\tilde{b}_1$ slowly reduces to 0 as a_2 gets closer to 1.

As a comparison, Figure 3.6(b) shows the opposite scenario where the increase in a_2 has smaller effect on ρ_{X,M_U} compared to ρ_{X,M_O} and ρ_{M_O,M_U}. That is, the increase in a_2 gets enlarged in ρ_{X,M_O} and ρ_{M_O,M_U} due to the large magnitudes of a_1 and k in ρ_{X,M_O} and ρ_{M_O,M_U}. As a result, X gains more and more comparative advantage in its competition with M_O. Under this scenario the increase in a_2 leads more explanatory power to the path $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ relative to the credit assigned to the pathway $X \rightarrow M_U \rightarrow Y$.

Figure 4.2 summarizes these patterns and interpretations about how a_2 affects the bias in estimating b_1 and c under different scenarios. Again, we see that the key regression model plays a significant role in helping us understand the mechanisms that generate the bias in $\tilde{b}_1$ and $\tilde{c}$ when we omit M_U.

4.4.3 How inconsistency changes with different levels of b_2.

For how b_2 affects the bias, we also presented two different scenarios in Chapter 3, exemplified by Figure 3.6(a) and 3.6(b). In both scenarios, the positive bias of $\tilde{b}_1$ becomes larger as b_2 gets closer to 1 while the level of b_2 has no effect on the magnitude of bias in a_1. The first two rows of Figure 4.1 can help us interpret these patterns by applying our framework.

First for $\tilde{a}_1$, we know that its bias comes from the explanatory power that belongs to the $X \rightarrow M_U \rightarrow M_O$ and there is no role of b_2 here. That is, the direct effect $M_U \rightarrow Y(b_2)$ is not directly connected with the direct effect $X \rightarrow M_O(a_1)$. Alternatively, we can think about this "no effect" by noticing that the parameter we manipulate b_2 occurs after the direct effect $X \rightarrow M_O(a_1)$ in the sequence of causality.

Figure 4.1 illustrates the bias in estimating b_1 is the product of β_2 and b_2. The magnitude of β_2 depends on the competition between M_O and X that happens in the key regression model represented in Figure 4.1. We can tell from the triangle of the key regression model that b_2 has no effect on this model. Therefore, the only way for b_2 to influence the bias in $\tilde{b}_1$ is to through b_2 but not β_2. This helps explain why the bias in $\tilde{b}_1$ is a linear relationship of b_2 in both Figure 3.6(a) and 3.6(b).

The distinction between Figure 3.6(a) and 3.6(b) is the pattern for the bias of c. In Figure

3.6(a), c gets overestimated and the positive bias keeps growing but in Figure 3.6(b) we always underestimate c and importantly, the negative bias gets more serious as b_2 increases. In fact, the fundamental reason underlying this distinction between two figures (3.6(a) and 3.6(b)) is the direction of bias while the effect of b_2 on the magnitude of bias follows the same pattern in both two figures. This ties back to the earlier discussion of the direction of $\tilde{c}$'s bias, which can be interpreted with our parameter framework again. In Figure 3.6(a) we have $a_1 = k = a_2 = 0.2$ but in Figure 3.6(b), $a_1(= 0.55)$ and $k(= 0.6)$ are much larger than $a_2(= 0.2)$. Correspondingly, we have $\rho_{X,M_U} > \rho_{X,M_O} \cdot \rho_{M_O,M_U}$ for Figure 3.6(a) and $\rho_{X,M_U} < \rho_{X,M_O} \cdot \rho_{M_O,M_U}$ for Figure 3.6(b). In other words, in Figure 3.6(b), the relationship between X and M_U gets suppressed or distorted once we include M_O to explain the outcome M_U but this suppression does not occur in Figure 3.6(a).

4.5 Using the Mechanism to Understand the Inconsistency of Indirect and Direct Effects When Omitting M_U

In the previous section, we applied our parameter framework to interpret how bias is generated in estimating each parameter we are interested in (i.e., a_1, b_1 and c). But we are also interested in the indirect effect via M_O, which is the product of a_1 and b_1. Importantly, as we will see later, this discussion can provide empirical researchers some ideas to consider how different potential candidates of M_U may impact our estimation of the indirect effect $a_1 b_1$ and direct effect c differently.

We start our discussion by getting the following equation for the bias in estimating $a_1 b_1$ (the specific indirect effect via M_O) based on our previous derivations.

$$\tilde{a}_1 \tilde{b}_1 - a_1 \cdot b_1 = a_2 \cdot k \cdot b_1 + b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) \tag{4.11}$$

Combing Equation 4.11 and 4.10, we can get:

$$\left[\tilde{a}_1 \tilde{b}_1 - a_1 \cdot b_1\right] + [\tilde{c} - c] = a_2 \cdot k \cdot b_1 + b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) + b_2 \cdot \beta_3 \tag{4.12}$$

In Equation 4.12, the two last components $b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) + b_2 \cdot \beta_3$ can be written as $a_2 \cdot b_2$ (see Appendix for more detailed proof). This allows us to get the following equation 4.13.

$$\left[\tilde{a}_1 \tilde{b}_1 - a_1 \cdot b_1\right] + [\tilde{c} - c] = a_2 \cdot k \cdot b_1 + a_2 \cdot b_2 \tag{4.13}$$

Equation 4.13 tells us an important fact: the bias in estimating the indirect effect $a_1 b_1$ and the bias in estimating the direct effect c adds up to the two pathways: $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$. Figure 4.3 summarizes this finding by comparing different pathways from X to Y by models: the true model with two mediators versus the model that omits M_U. The left column shows four causal pathways in the true model from X to Y. As a comparison, there are only two causal pathways in the right column to account for the total effect from X to Y. Because the total

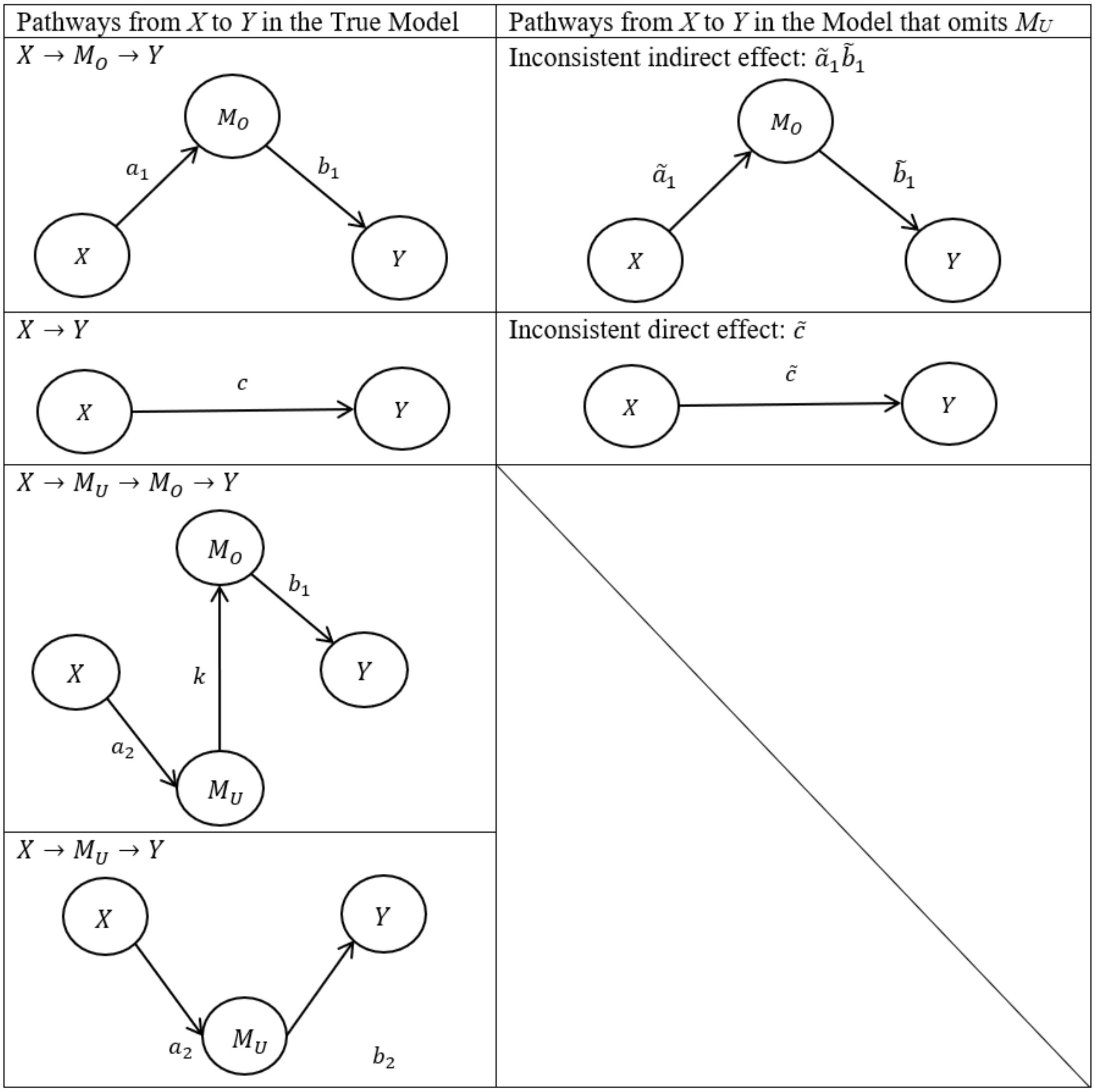

Pathways from X to Y in the True Model
Pathways from X to Y in the Model that omits M_U
$X \rightarrow M_O \rightarrow Y$
M_O
a_1
b_1
X
Y
Inconsistent indirect effect: $\tilde{a}_1 \tilde{b}_1$
M_O
$\tilde{a}_1$
$\tilde{b}_1$
X
Y
$X \rightarrow Y$
c
X
Y
Inconsistent direct effect: $\tilde{c}$
$\tilde{c}$
X
Y
$X \rightarrow M_U \rightarrow M_O \rightarrow Y$
M_O
b_1
X
k
Y
a_2
M_U
$X \rightarrow M_U \rightarrow Y$
X
Y
a_2
M_U
b_2

Figure 4.3: Different causal pathways from X to Y by models: the true model with two mediators versus the model that omits M_U.

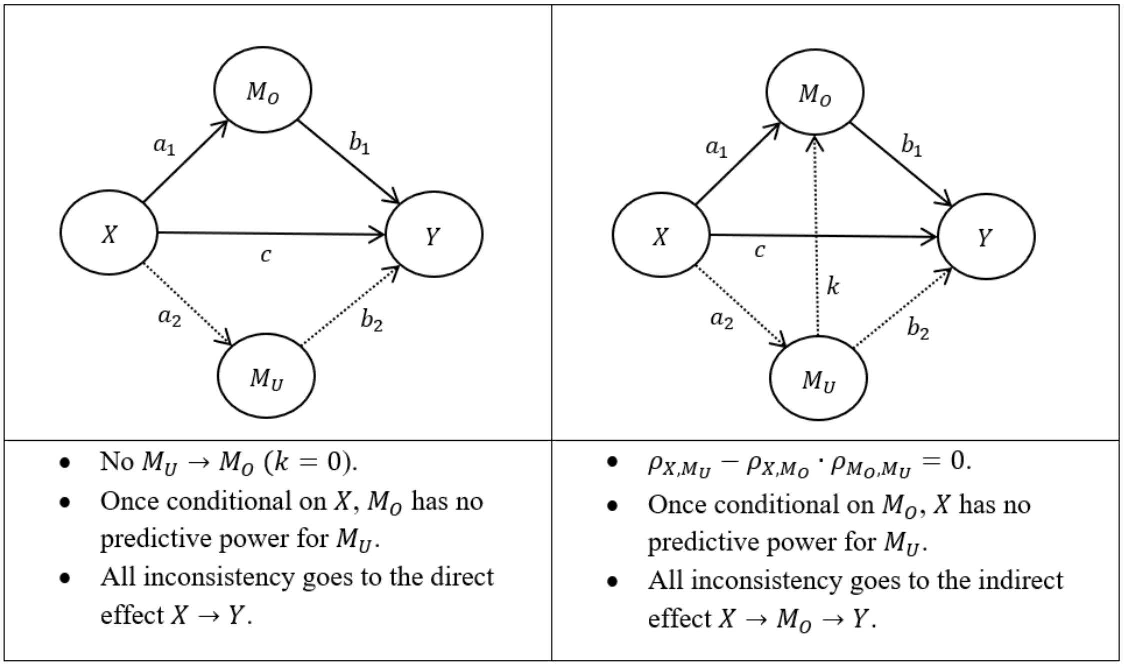
effect from X to Y is fixed, the effect via the two missing pathways, namely $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$, must be assigned to the two remaining pathways when M_U is excluded. In other words, the explanatory power via $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$ either goes to the indirect effect $\tilde{a}_1\tilde{b}_1$ or the direct effect $\tilde{c}$. When more effect is allocated to the indirect effect, then omitting M_U results in more biased estimate of the indirect effect and less biased estimate of the direct effect. The same vice versa: if more effect is allocated to the direct effect, then omitting M_U generates more biased direct effect and less biased indirect effect.

What factors affect how the effect via $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$ is allocated to the bias in the estimated indirect effect or the bias in the estimated direct effect? The answer to these questions goes back to our parameter mechanisms, more specifically, the competition between X and M_O in the key regression model. We know the bias either goes to the estimated indirect effect or the estimated direct effect, and it is easier to focus on the direct effect. As we discussed before, the bias in estimating the direct effect c is $b_2 \cdot \beta_3$. b_2 does not even show up in the key regression model or the competition between X and M_O. Then as the exposure variable X gets stronger at predicting the unobserved mediator M_U, compared to the observed mediator M_O, more bias will be allocated to the estimated direct effect and less bias will be allocated to the estimated indirect effect. In other words, when b_2 is fixed at a certain level (not zero), if the unobserved mediator M_U becomes more correlated to X and less correlated to M_O, then omitting M_U generates more biased direct effect and less biased indirect effect.

It is important to note that this "more" or "less" is not a comparison between the amount of bias between the estimated indirect effect and the estimated direct effect. Instead, it is a trend in the change of the bias of either the indirect effect itself or the direct effect itself. Assume an example where 10% of the total bias (i.e., the effects via $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$) is allocated to the estimated direct effect and other 90% of the total bias is allocated to the estimated indirect effect. Now as the unobserved mediator M_U gets more correlated with X and less correlated with M_O, maybe 20% of the bias is allocated to the estimated direct effect and the other 80% is allocated to the direct effect. But still more bias gets assigned to the indirect effect in this case

(80% versus 20%).

With this understanding, now we can better understand two special situations, as summarized in Figure 4.4. In the first context, the true underlying process is a parallel mediation model. M_U does not have a direct effect on M_O and consequently, M_U and M_O are independent once conditional on X. That is, X wins all possible explanatory power in the competition with M_O. Once we have X, M_O has no predictive power for M_U. As such, the direct effect is mostly overestimated, accounting for all the explanatory credit that belongs to the omitted $X \rightarrow M_U \rightarrow Y$. In contrast, the indirect effect $X \rightarrow M_O \rightarrow Y$ is not affected at all by the omitted mediator (i.e., the estimated indirect effect is consistent).



- No $M_U \rightarrow M_O$ ($k = 0$).
- Once conditional on X, M_O has no predictive power for M_U.
- All inconsistency goes to the direct effect $X \rightarrow Y$.

- $\rho_{X,M_U} - \rho_{X,M_O} \cdot \rho_{M_O,M_U} = 0$.
- Once conditional on M_O, X has no predictive power for M_U.
- All inconsistency goes to the indirect effect $X \rightarrow M_O \rightarrow Y$.

Figure 4.4: Two special situations where all inconsistency goes to the direct effect $X \rightarrow Y$ or all inconsistency goes to the indirect effect $X \rightarrow M_O \rightarrow Y$.

In the second context, we have $\rho_{X,M_U} - \rho_{X,M_O} \cdot \rho_{M_O,M_U} = 0$. This means X has no predictive power for M_U once conditional on M_O. Then M_O wins all the explanatory credit possible in its competition with X to predict M_U. As such, all the bias is allocated to the estimated indirect effect $X \rightarrow M_O \rightarrow Y$ and the direct effect $X \rightarrow Y$ is consistent.

Note the key in the second situation is: once conditional on M_O, X has no predictive power for M_U. This does not happen when there is no effect from X to $M_U (a_2 = 0)$. That is, when M_U is only a confounder for $M_O \rightarrow Y$ but not a mediator, even conditional on M_O, X can still have a predictive effect on M_U. Another important note here is that in both two situations in Figure 4.4, there is a path of $M_U \rightarrow Y$ (i.e., $b_2 \neq 0$). When $b_2 = 0$, only $\tilde{a}_1$ is biased. As such, only the estimated indirect effect is biased, and direct effect is unbiased. The serial mediation model is a special situation of this where $a_1 = b_2 = 0$.

4.6 Applying the Mechanism to Understand the Inconsistency in Longitudinal Designs When Omitting M_U

Recent research has suggested longitudinal designs to test mediation because cross-sectional examination of mediation can generate biased estimates and longitudinal designs provide more rigorous inference for mediation effects (e.g., Maxwell & Cole, 2007; Maxwell et al., 2011; Mitchell & Maxwell, 2013). Yet there may be post-treatment confounders M_U invalidating mediation effects even in longitudinal designs. This section will present how our parameter framework can help us understand the bias generation in a longitudinal design. More specifically, this section will focus on one autoregressive model presented by Maxwell et al. (2011). Following Figure 4.5 presents the path diagram for this model and the formal equations are followed.

This model can be written as follows:

$$X_{i,t+1} = xX_{i,t} + \varepsilon_{X_{i,t+1}} \tag{4.14}$$

$$M_{i,t+1} = mM_{i,t} + aX_{i,t} + \varepsilon_{M_{i,t+1}} \tag{4.15}$$

$$Y_{i,t+2} = yY_{i,t+1} + bM_{i,t+1} + cX_{i,t} + \varepsilon_{Y_{i,t+2}} \tag{4.16}$$

where $X_{i,t+1}$ is the treatment status for individual i at time $t + 1$, $X_{i,t}$ is the treatment status for individual i at time t, $M_{i,t+1}$ is the value for individual i on mediator M at time $t + 1$, $M_{i,t}$ is the value for individual i on mediator M at time t, $Y_{i,t+2}$ is the value for individual i on outcome Y at time $t + 2$, $Y_{i,t+1}$ is the value for individual i on outcome Y at time $t + 1$, and similarly, $Y_{i,t}$ is the value for individual i on outcome Y at time t. We can also tell that the direct effect in this model is the product of a and b while the indirect effect is c. We assumed that this longitudinal model

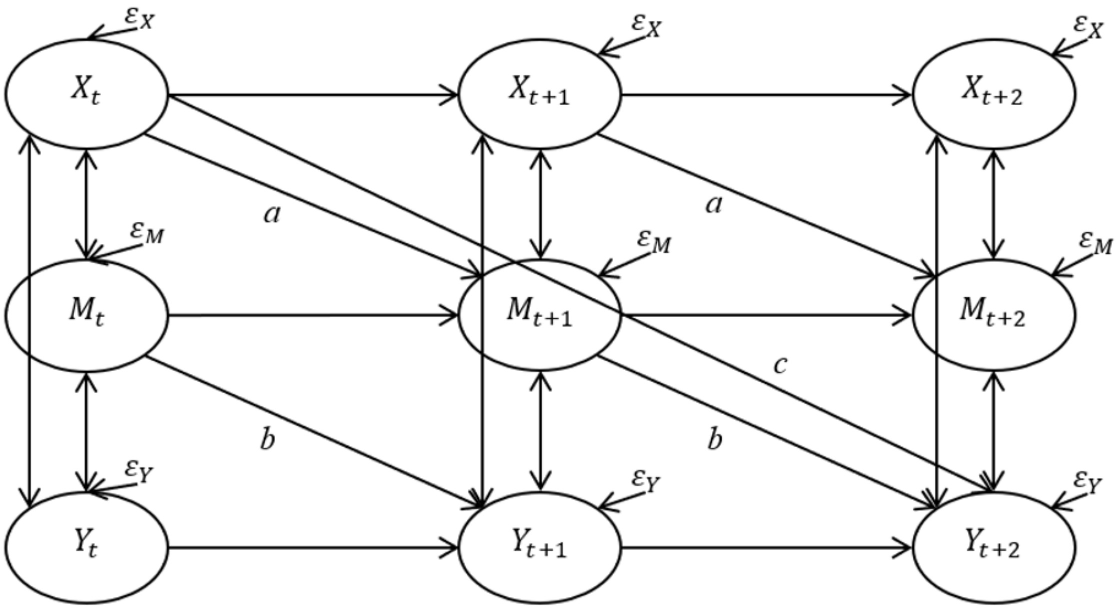


Figure 4.5: Longitudinal mediation model with two unit lag for direct effect of X on Y.

satisfies the stationarity and equilibrium. This means that the causal relationships among variables and the within-wave correlations are unchanged over time (the variance-covariance matrix among $X_{i,t}$, $Y_{i,t}$ and $M_{i,t}$ is time invariant).

We then introduced an unobserved mediator M_U into this model and allowed this mediator to have an effect on the observed mediator. To make a clear distinction between these two mediators and make the notation consistent with our cross-sectional design, we note the observed mediator of interest as M_O. The path diagram is presented in the following Figure 4.6.

This model can be written formally as follows:

$$X_{i,t+1} = xX_{i,t} + \varepsilon_{X_{i,t+1}} \tag{4.17}$$

$$M_{Oi,t+1} = m_1 M_{Oi,t} + kM_{Ui,t} + a_1 X_{i,t} + \varepsilon_{M_{Oi,t+1}} \tag{4.18}$$

$$M_{Ui,t+1} = m_2 M_{Ui,t} + a_2 X_{i,t} + \varepsilon_{M_{Ui,t+1}} \tag{4.19}$$

$$Y_{i,t+2} = yY_{i,t+1} + b_1 M_{Oi,t+1} + b_2 M_{Ui,t+1} + cX_{i,t} + \varepsilon_{Y_{i,t+2}} \tag{4.20}$$

where the notation is almost the same as before. The only distinction is that another mediator M_U is introduced so we use M_O to represent the observed mediator. Similarly, the indirect effect

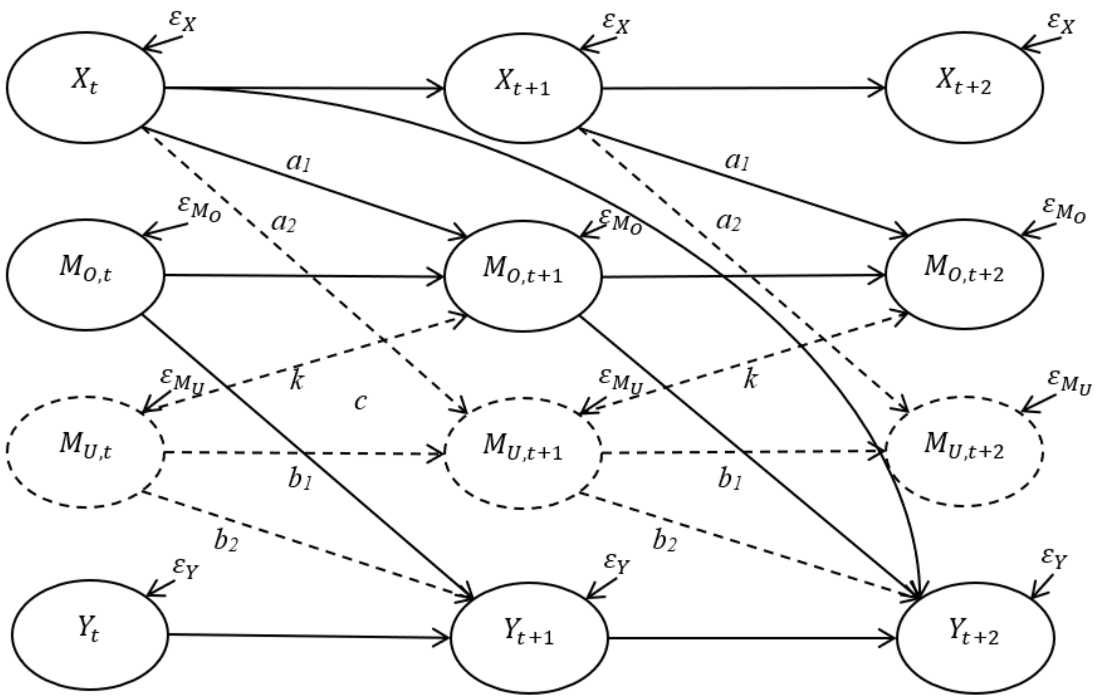


Figure 4.6: Longitudinal mediation model with unobserved latent mediator M_U.

associated with the observed mediator is the product of a_1 and b_1. Additionally, the autoregressive parameters for the two mediators are noted as m_1 and m_2 and k represents the causal effects of $M_{Ui,t}$ on $M_{Oi,t+1}$. As before, we assumed stationarity and equilibrium, indicating that the causal relationship and the correlation matrix among X, M_O, M_U, and Y are time-invariant. Note in this longitudinal model, the unobserved mediator M_U serves as a lagged confounding, but not simultaneous confounding. That is, the longitudinal design only has an effect $M_{U,t} \rightarrow M_{O,t+1}$ but not $M_{U,t+1} \rightarrow M_{O,t+1}$.

Now we apply the same approach LIE to derive the inconsistency in estimating a_1, b_1, and c when omitting M_U. The procedure is essentially the same. We first write our true models as:

$$E\left(X_{t+1}|X_t\right) = x \cdot X_t \tag{4.21}$$

$$E\left(M_{O,t+1}|M_{U,t}, M_{O,t}, X_t\right) = m_1 \cdot M_{O,t} + k \cdot M_{U,t} + a_1 \cdot X_t \tag{4.22}$$

$$E\left(M_{U,t+1}|X_t, M_{U,t}\right) = m_2 \cdot M_{U,t} + a_2 \cdot X_t \tag{4.23}$$

$$E\left(Y_{t+2}|M_{O,t+1},\ M_{U,t+1}, X_t\right) = y \cdot Y_{t+1} + b_1 \cdot M_{O,t+1} + b_2 \cdot M_{U,t+1} + c \cdot X_t \tag{4.24}$$

By LIE, we can further write above Equation 4.22 and 4.24 to see what happens when M_U gets excluded:

$$\begin{aligned} E\left(M_{O,t+1}|M_{O,t}, X_t\right) &= E\left[E\left(M_{O,t+1}|M_{U,t}, M_{O,t}, X_t\right)|M_{O,t}, X_t\right] \\ &= m_1 \cdot M_{O,t} + k \cdot E\left(M_{U,t}|X_t, M_{O,t}\right) + a_1 \cdot X_t \qquad (4.25) \\ E\left(Y_{t+2}|M_{O,t+1},\ X_t\right) &= E\left[E\left(Y_{t+2}|M_{O,t+1},\ M_{U,t+1}, X_t\right)|M_{O,t+1},\ X_t\right] \\ &= y \cdot Y_{t+1} + b_1 \cdot M_{O,t+1} + b_2 \cdot E\left(M_{U,t+1}|X_t, M_{O,t+1}, Y_{t+1}\right) + c \cdot X_t \qquad (4.26) \end{aligned}$$

Now write:

$$E\left(M_{U,t}|X_t, M_{O,t}\right) = \gamma_1 \cdot M_{O,t} + \gamma_2 \cdot X_t \qquad (4.27)$$

$$E\left(M_{U,t+1}|X_t, M_{O,t+1}, Y_{t+1}\right) = \beta_1 \cdot Y_{t+1} + \beta_2 \cdot M_{O,t+1} + \beta_3 \cdot X_t \qquad (4.28)$$

Plugging these two equations back to Equations 4.25 and 4.26, we can get formulas for inconsistency in estimating a_1, b_1 and c, as follows, where $\tilde{a}_1$, $\tilde{b}_1$, and $\tilde{c}$ are estimated effect of a_1, b_1 and c when M_U is excluded (again, assuming we have population level data, $n \to \infty$).

$$\tilde{a}_1 - a_1 = k \cdot \gamma_2 \qquad (4.29)$$

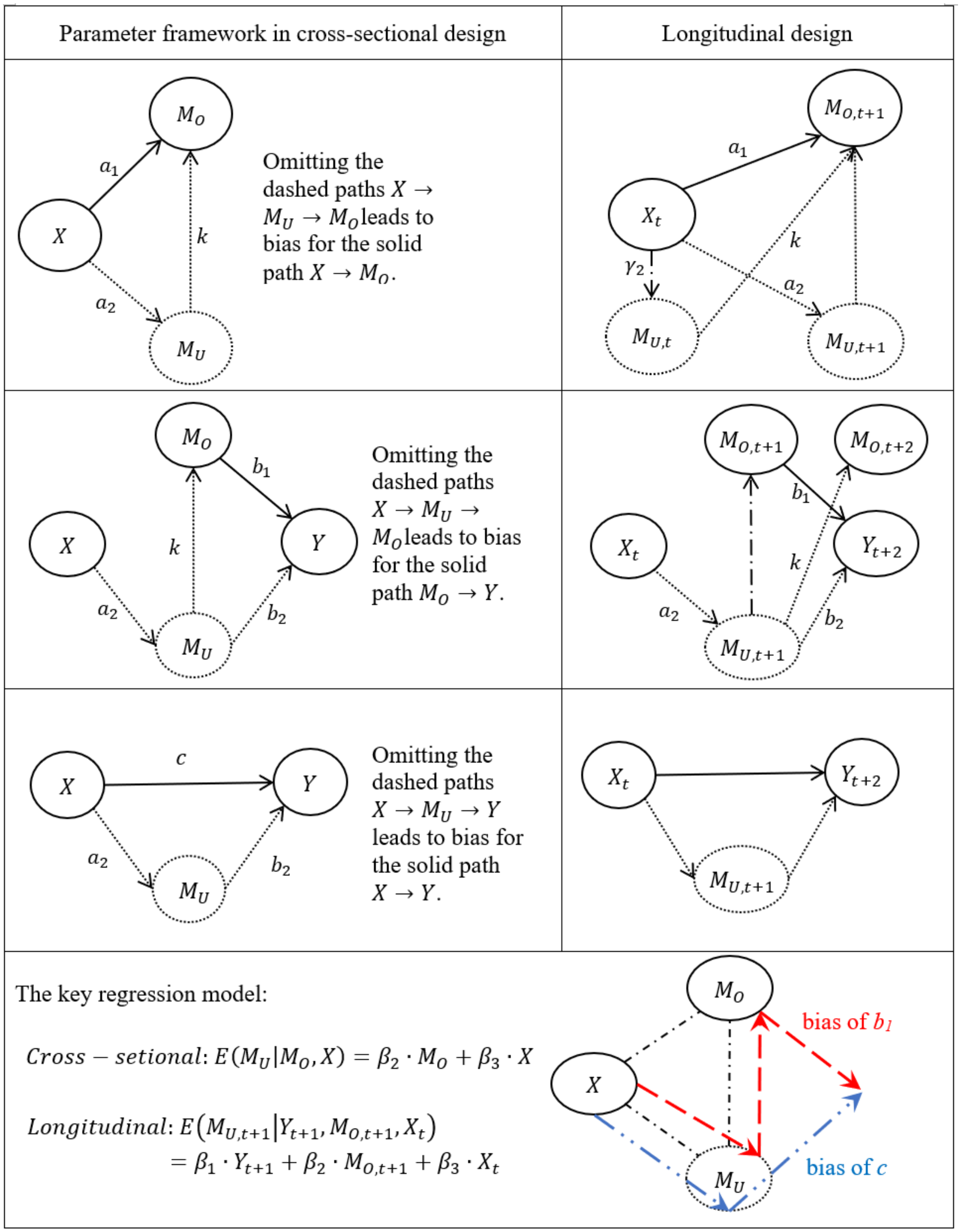
$$\tilde{b}_1 - b_1 = b_2 \cdot \beta_2 \qquad (4.30)$$

$$\tilde{c} - c = b_2 \cdot \beta_3 \qquad (4.31)$$

Importantly, as implied by Equation 4.27 and 4.28, γ_2, β_2 and β_3 are three regression coefficients. Specifically, γ_2 is the regression coefficient of X_t when we regress $M_{U,t}$ on X_t and $M_{O,t}$. β_2 and β_3 are regression coefficients of $M_{O,t+1}$ and X_t, respectively, when regressing $M_{U,t+1}$ on Y_{t+1}, $M_{O,t+1}$ and X_t.

Figure 31 presents how our parameter mechanism can be extended to this autoregressive longitudinal mediation model to understand how bias (more precisely, inconsistency) is generated when omitting M_U. As we will see, the major distinction between the cross-sectional and longitudinal design is the control of prior X, M_U, M_O and Y.

We start our consideration with $\tilde{a}_1$. In the cross-sectional design, it is evident to see from the path diagram that the bias in $\tilde{a}_1$ comes from the omitted causal pathway $X \to M_U \to M_O$. $\tilde{a}_1$



Omitting the dashed paths $X \rightarrow M_U \rightarrow M_O$ leads to bias for the solid path $X \rightarrow M_O$.

Omitting the dashed paths $X \rightarrow M_U \rightarrow M_O$ leads to bias for the solid path $M_O \rightarrow Y$.

Omitting the dashed paths $X \rightarrow M_U \rightarrow Y$ leads to bias for the solid path $X \rightarrow Y$.

The key regression model:

$$Cross - setional: E(M_U|M_O, X) = \beta_2 \cdot M_O + \beta_3 \cdot X$$

$$Longitudinal: E\left(M_{U,t+1}\middle|Y_{t+1}, M_{O,t+1}, X_t\right) = \beta_1 \cdot Y_{t+1} + \beta_2 \cdot M_{O,t+1} + \beta_3 \cdot X_t$$

Figure 4.7: Parameter framework extended to longitudinal designs.

measures the effect $X \rightarrow M_O$ plus the effect $X \rightarrow M_U \rightarrow M_O$. Now in the longitudinal scenario, the mechanism is essentially the same: $\tilde{a}_1$ is biased because it accounts for the effect $X_t \rightarrow M_{U,t} \rightarrow M_{O,t+1}$ that should be attributed to $M_{U,t}$ as a mediator. The effect of $M_{U,t} \rightarrow M_{O,t+1}$ is captured by k in the formula but we do not have a parameter for $X_t \rightarrow M_{U,t}$. We only have a_2 to represent $X_t \rightarrow M_{U,t+1}$ in the longitudinal scenario. This is why in the formula we have γ_2 rather than a_2. Loosely speaking, γ_2 captures what the effect of $X_t \rightarrow M_{U,t+1}(a_2)$ projects on $X_t \rightarrow M_{U,t}$. However, γ_2 is not equivalent to the correlation between X_t and $M_{U,t}$ because both X_t and $M_{U,t}$ are also determined by their prior values, namely X_{t-1} and $M_{U,t-1}$. That explains why γ_2 is the regression coefficient of X_t when we regress $M_{U,t}$ on X_t and $M_{O,t}$. The presence of $M_{O,t}$ in this regression is exactly controlling for X_{t-1} and $M_{U,t-1}$, because our model implies that $M_{O,t}$ is determined by X_{t-1}, $M_{U,t-1}$ and its own prior value $M_{O,t-1}$.

The stories for $\tilde{b}_1$ and $\tilde{c}$ are also essentially the same with the cross-sectional design. In the longitudinal case we still observe that the bias in estimating b_1 and c are both a weighted version of $b_2(M_{U,t} \rightarrow Y_{t+1})$. The two weights for the bias in estimating b_1 and c are still two partial regression coefficients from one regression model. The last row of Figure 4.7 presents a comparison between the key regression model in the cross-sectional desgin versus that in the longitudinal design. In both situations, we use X and M_O to predict M_U, and the weight for the bias in estimating b_1 is the coefficient of M_O and the weight for the bias in estimating c is the coefficient of X. This similarity illustrates the competition between X and M_O is still present in the longitudinal situation: the unique contribution made by M_O to explain M_U composes the weight before b_2 to generate the bias in $\tilde{b}_1$, and the unique contribution attributed by X to explain M_U becomes the weight of b_2 to form the bias in $\tilde{c}$. The argument about M_U serving as a confounder in the bias of $\tilde{b}_1$ and as a mediator in the bias of $\tilde{c}$ is also the same in the longitudinal design.

What is different in the longitudinal design is that now we need to consider time points and controlling for prior values. In the cross-sectional design, we argue that the competition between X and M_O comes from the fact: pathway $X \rightarrow M_U$ splits into two pathways $X \rightarrow M_U \rightarrow M_O \rightarrow Y$ and $X \rightarrow M_U \rightarrow Y$. Now the longitudinal story is a little different: it is the pathway of $X_t \rightarrow M_{U,t+1}$

that splits into $X_t \to M_{U,t+1} \to M_{O,t+1} \to Y_{t+2}$ and $X_t \to M_{U,t+1} \to Y_{t+2}$. This first part (i.e., $X_t \to M_{U,t+1} \to M_{O,t+1} \to Y_{t+2}$) passes through the b_1 pathway and contributes to the bias in $\tilde{b}_1$. The second part (i.e., $X_t \to M_{U,t+1} \to Y_{t+2}$) picks up a mediation effect transmitted through $M_{U,t+1}$ and generates the bias in $\tilde{c}$. This helps us interpret why the key regression model in the longitudinal design is regressing $M_{U,t+1}$ on $M_{O,t+1}$ and X_t and why the coefficient of $M_{O,t+1}$ is responsible for the bias in $\tilde{b}_1$ while the coefficient of X_t is responsible for the bias in $\tilde{c}$. Note the longitudinal design does not really have the effect of $M_{U,t+1} \to M_{O,t+1}$ since $M_{U,t+1}$ and $M_{O,t+1}$ are at the same time point. But one can loosely interpret this as a projection of the effect $M_{U,t+1} \to M_{O,t+2}$, just as previously how we interpret γ_2 as a projection of the effect $X_t \to M_{U,t+1}$.

Now we only need to explain why Y_{t+1} shows up as a control variable in the key regression model in longitudinal design. This is similar to why $M_{O,t}$ is controlled to generate γ_2. Here Y_{t+1} is present as a control for prior values that can have an effect on $M_{U,t+1}$, $M_{O,t+1}$ and X_t, namely $M_{U,t}$, $M_{O,t}$ and X_{t-1}. This ties back to the argument about the fundamental difference between the cross-sectional and longitudinal design: as Maxwell and Cole (2007) argued, what is missing in cross-sectional examination of mediation is the failure to capture the autoregressive effects.

To summarize, although the longitudinal design to examine mediation is much more complicated, the interpretation of how bias (more precisely, inconsistency) is generated with omitting M_U is essentially the same as the cross-sectional design. The parameter framework depicted in this chapter allows us to grasp an intuitive path-based understanding of how omitting M_U generates bias in both cross-sectional and longitudinal designs.

4.7 Discussion

This chapter leverages the parameter framework to explain how inconsistency is generated when an unobserved mediator is omitted. Briefly speaking, the inconsistency of $\tilde{a}_1$ is due to omitting $X \to M_U \to M_O$ and thus part of the direct effect $X \to M_O$ should be allocated to this omitted indirect effect $X \to M_U \to M_O$ via M_U. The inconsistency of $\tilde{b}_1$ comes from the omission of $X \to M_U \to M_O \to Y$ and the inconsistency of $\tilde{c}$ comes from the omission of $X \to M_U \to Y$.

Importantly, we can consider the fact that the inconsistency of $\tilde{b}_1$ and $\tilde{c}$ both originates from $X \rightarrow M_U$ as competition between X and M_O to predict M_U in a linear regression framework. As the exposure variable X gets stronger at predicting the unobserved mediator M_U, compared to the observed mediator M_O, more bias will be assigned to $\tilde{c}$ and less bias will be assigned to $\tilde{b}_1$.

Applying this framework, we can better understand how each M_U-related parameter affects the inconsistency, how inconsistency is allocated to either the direct effect from X to Y or the indirect effect via M_O, as well as inconsistency in some special situations including the parallel and serial mediation models. Additionally, I showed that the inconsistency underlying a longitudinal design is essentially the same, except that prior values are controlled in the longitudinal design.

4.8 Limitations and Future Directions

There are at least two limitations of the current chapter that suggest avenues for future research. First, this chapter focus on understanding how inconsistency is generated when the alternative mediator is omitted. Future studies can look at how this understanding can be applied in a more practical perspective. For example, can we bound the inconsistency for the indirect ($\tilde{a}_1\tilde{b}_1$) and direct ($\tilde{c}$) effects if we have some ideas about the correlation between the unobserved mediator M_U and the intervention X, and the correlation between M_U and the mediator of interest M_O? Second, it would be valuable if future studies can develop a sensitivity approach for an unobserved post-treatment confounder in a longitudinal design, as what we do in Chapter 3 for the cross-sectional model.

DISCUSSION

As mentioned in the introduction, this dissertation is centered on causal inference regarding evaluation of an intervention, from whether an intervention works to why it works, accounting for the social and dynamic contexts in which interventions are implemented. The first two chapters focus on whether an intervention works by proposing a non-parametric case replacement framework to quantify strength of evidence for inferences in multisite randomized control trials and value-added measures for teacher effectiveness. The last two chapters focus on why an intervention works by studying post-treatment confounders in mediating processes.

From a practical perspective, the four chapters represent the effort to refine policy analysis so that policies regarding the allocation of educational resources can be better informed. As the first step, we are interested in the total average intervention effect. But summarizing an intervention with only one estimated effect can be misleading, especially in the context of multisite randomized control trials. For example, presence of heterogeneity in multisite randomized control trials emphasizes the importance of considering local contextual effects and causal mechanisms that can help explain why an intervention works in some sites but not others. Identifying important mediating pathways may point out an alternative intervention option, especially when the mediating pathway can explain a large proportion of the causal effect and it is easier or more cost-efficient to manipulate the mediator directly. Further, considering alternative mediators as potential confounders can help researchers and policy makers evaluate the robustness of the inference regarding the identified mediator of interest so that the reallocation of educational resources regarding the mediator of interest can be made after comprehensive evaluation against potential costs and other alternatives.

There are other ways that mediation studies can inform policy manipulations. For example, sometimes mediation analysis may present us with two mediating pathways with opposite directions that cancel out each other and lead to a zero total effect (i.e., average treatment effect), which provides another example for considering manipulating one mediator instead to achieve the expected changes in the outcome. Under other scenarios, a mediator can also be a side effect of interest that inheres in the intervention. Then knowing the presence of this mediator can allow us to figure out ways to minimize the negative side effects.

Problems of causal inference can be conceptualized in terms of omitted variables, or in terms of sampling bias. Both mediators and confounders are third variables that influence the association between the predictor of interest and the outcome, and they are statistically identical (e.g., MacKinnon et al., 2000). In Chapters 1 and 2, the presence of spillover effects violates SUTVA and generates bias via playing a similar role as a confounder: by associating with both experiment conditions and outcome measures. Importantly, the association between spillover effects and experiment condition implies the treatment and control group individuals experience different levels of spillover effects. For example, positive spillover effects within the treatment group, or negative spillover effects within the control group, due to non-random assignment of contributors and spoilers to treatment and control conditions, can bias the treatment effect estimate negatively (assuming a positive treatment effect). Sometimes capacity constraints can also lead to negative spillover effects within treatment group (e.g., Maroulis, 2016), leading to downward bias in the treatment effect estimation. Alternatively, the treatment condition triggers positive spillover effects from treatment individuals to control individuals, or the control condition triggers negative spillover effects from control group to treatment group, both of which may cause downward bias in estimating the treatment effect (again, assuming a positive treatment effect). In other situations, affecting individuals' interactions to introduce spillover effects can be one *mediating* process that explains why the intervention works. For example, if reducing class size improves students' learning through maximizing the learning and teaching among students, then the peer effect becomes a *mediator* that helps explain how smaller classes boost students' performance.

Furthermore, constant effects through simple mechanisms to independent individuals rarely occur in education research (Frank, Saw, & Xu, 2016). Hong (2015) conceptualizes causal inference regarding moderation, mediation and spillover as weighting issues in a sampling framework. Relatedly, the case replacement approach proposed in the first two chapters for spillover and heterogeneity applies the feature in the counterfactual framework that recast potential sources of bias in terms of missing data (Frank et al., 2013). Both replacing observed cases with unobserved cases and adjusting weights in observed cases are rooted in a sampling framework. Chapters 3 and 4 study mediation within a parametric framework, but they also contribute to existing literature that applies a non-parametric approach (e.g., Hong et al., 2018) by taking on a path coefficient perspective and digging into the effects of a confounding mediator on each path-coefficient estimate. We argue that parametric and non-parametric approaches complement each other so that researchers can better understand spillover, heterogeneity, and mediation in education research.

Finally, it is important to note that sensitivity analysis cannot exclude bias, regardless of what type of approach one uses, either traditional approaches that draw on familiar quantities such as correlations or percentage of variance explained, or the non-parametric case replacement approach proposed in Chapters 1 and 2. Before applying sensitivity analysis, one should make sure that best considerations have been given to research design, model specification, and choice of estimation approach. Sensitivity analysis cannot substitute any of these crucial steps, but rather provides a discourse for researchers to communicate with all potential stakeholders regarding the strength of evidence, after all those efforts are made to remove as much bias as possible.

APPENDIX

DERIVATION NOTE FOR UNOBSERVED MEDIATOR IN A CROSS-SECTIONAL DESIGN

Introduction

This Appendix derives the inconsistency an unobserved mediator M_U may bring to the estimation of the indirect and direct effects in a cross-sectional design. The key approach applied in this derivation is the Law of Iterated Expectations (LIE). I will introduce the true model in Part 1 and then write the true model in terms of conditional mean (Part 2). After that, I derive several correlations that are useful in later derivation (Part 3). In Part 4, I apply LIE to derive the inconsistency as a function of parameters that have been derived in Part 3. I work out the formulas for inconsistency as a function of only correlations in Part 5 and the formulas for percent of inconsistency in Part 6. Part 7 includes the derivation of the error variances for the purpose of simulation (to generate standardized variables) and also constraints of parameters. Part 8 and 9 discuss the directions of inconsistency (Part 8) and how the inconsistency changes with parameters (Part 9). The last part includes some derivation to decompose the total inconsistency into two omitted pathways.

One crucial assumptions made in this derivation is that all variables are standardized.

True model

Note: the causal relationship between two mediators is: M_U causes M_O.

$$M_O = k \cdot M_U + a_1 \cdot X + \epsilon_{M_O} \tag{32a}$$

$$M_U = a_2 \cdot X + \epsilon_{M_U} \tag{32b}$$

$$Y = b_1 \cdot M_O + b_2 \cdot M_U + c \cdot X + \epsilon_Y \tag{32c}$$

True model in terms of conditional mean

Note: the assumptions underlying the model are actually stronger than this. But the assumptions listed here are enough for the purpose of derivation in this Appendix.

$$E(M_O|M_U, X) = k \cdot M_U + a_1 \cdot X \tag{33a}$$

$$E(M_U|X) = a_2 \cdot X \tag{33b}$$

$$E(Y|M_O, M_U, X) = b_1 \cdot M_O + b_2 \cdot M_U + c \cdot X \tag{33c}$$

Correlations

Correlation between X and M_U

From Eq.32b:

$$cov(X, M_U) = a_2 \cdot var(X)$$

Assuming all variables are standardized:

$$\rho_{X,M_U} = a_2$$

Correlation between X and M_O

From Eq.32a:

$$cov(X, M_O) = k \cdot cov(X, M_U) + a_1 \cdot var(X)$$

Assuming all variables are standardized:

$$\rho_{X,M_O} = ka_2 + a_1$$

Correlation between M_O and M_U

From Eq.32a:

$$cov(M_O, M_U) = k \cdot var(M_U) + a_1 \cdot cov(X, M_U)$$

Assuming all variables are standardized:

$$\rho_{M_O,M_U} = k + a_1 a_2$$

Correlation between Y and M_U

From Eq.32c:

$$cov(M_U, Y) = b_1 \cdot cov(M_U, M_O) + b_2 \cdot var(M_U) + c \cdot cov(M_U, X)$$

Assuming all variables are standardized:

$$\rho_{Y,M_U} = b_1 k + b_1 a_1 a_2 + b_2 + ca_2$$

To summarize,

$$\rho_{X,M_U} = a_2 \tag{34a}$$

$$\rho_{X,M_O} = ka_2 + a_1 \tag{34b}$$

$$\rho_{M_O,M_U} = k + a_1 a_2 \tag{34c}$$

$$\rho_{Y,M_U} = b_1 k + b_1 a_1 a_2 + b_2 + ca_2 \tag{34d}$$

Inconsistency if omitting M_U

By Law of Iterated Expectation (LIE), the right model that excludes M_U can be written as follows:

$$\begin{aligned} E(M_O|X) &= E[E(M_O|M_U, X)|X] \\ &= k \cdot E(M_U|X) + a_1 \cdot X \qquad (35a) \\ E(Y|M_O, X) &= E[E(Y|M_O, M_U, X)|M_O, X] \\ &= b_1 \cdot M_O + b_2 \cdot E(M_U|M_O, X) + c \cdot X \qquad (35b) \end{aligned}$$

Write:

$$E(M_U|X) = \beta_1 \cdot X \tag{36a}$$

$$E(M_U|M_O, X) = \beta_2 \cdot M_O + \beta_3 \cdot X \tag{36b}$$

Then we get:

$$E(M_O|X) = k\beta_1 \cdot X + a_1 \cdot X \tag{37a}$$

$$E(Y|M_O, X) = b_1 \cdot M_O + b_2 \cdot \beta_2 \cdot M_O + b_2 \cdot \beta_3 \cdot X + c \cdot X \tag{37b}$$

Therefore, we can see that if we omitting M_U, we can get inconsistent estimates.

$$\tilde{a}_1 = a_1 + k\beta_1 \tag{38a}$$

$$\tilde{b}_1 = b_1 + b_2\beta_2 \tag{38b}$$

$$\tilde{c} = c + b_2\beta_3 \tag{38c}$$

The inconsistency for $\tilde{a}_1$ would be: $k \cdot \beta_1$;

The inconsistency for $\tilde{b}_1$ would be: $b_2 \cdot \beta_2$;

The inconsistency for indirect effect $\tilde{a}_1\tilde{b}_1$ would be: $(a_1 + k\beta_1) \cdot (b_1 + b_2\beta_2) - a_1b_1$;

The inconsistency for $\tilde{c}$ would be: $b_2 \cdot \beta_3$.

Following are derivations for $\beta_1, \beta_2, \beta_3$

From Eq.36a: β_1 is the regression coefficient of X when we regress M_U on X (at the population level). Therefore,

$$\beta_1 = \rho_{X,M_U} = a_2$$

Similarly, from Eq.36b: β_2 is the regression coefficient of M_O when we regress M_U on M_O and X; β_3 is the regression coefficient of X when we regress M_U on M_O and X (at the population level).

$$\beta_2 = \frac{\rho_{M_O,M_U} - \rho_{X,M_U} \cdot \rho_{X,M_O}}{1 - \rho^2_{X,M_O}} \tag{39a}$$

$$\beta_3 = \frac{\rho_{X,M_U} - \rho_{M_O,M_U} \cdot \rho_{X,M_O}}{1 - \rho^2_{X,M_O}} \tag{39b}$$

$$\tag{39c}$$

where all elements have been derived in Eq.34.

Note: β_2 and β_3 are derived based on formula of regression coefficients as follows (I derived both of these through the well-known $(X'X)^{-1}X'Y$ and some linear algebra).

When we regress Y on $X1$ and $X2$, the coefficient of $X1$ is (in terms of correlations):

$$\frac{\rho_{Y,X1} - \rho_{Y,X2} \cdot \rho_{X1,X2}}{1 - \rho_{X1,X2}^2} \tag{40}$$

To summarize, we can get following estimates if we omit M_U:

$$\tilde{a}_1 = a_1 + k \cdot a_2 \tag{41a}$$

$$\tilde{b}_1 = b_1 + b_2 \cdot k \cdot \frac{1 - a_2^2}{1 - (k \cdot a_2 + a_1)^2} \tag{41b}$$

$$\tilde{c} = c + b_2 \cdot \frac{a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)}{1 - (k \cdot a_2 + a_1)^2} \tag{41c}$$

We can also get:

$$\tilde{a}_1 \cdot \tilde{b}_1 = a_1 \cdot b_1 + k \cdot a_2 \cdot b_1 + (a_1 + k \cdot a_2) \cdot \frac{b_2 \cdot k \cdot (1 - a_2^2)}{1 - (k \cdot a_2 + a_1)^2} \tag{42}$$

Inconsistency as a function of correlations only

This section aims to write the inconsistent estimators as a function only of correlations. To achieve this, we start from Eq.38a, Eq.38b, Eq.38c and formulas for β_1, β_2 and β_3. From these equations, we can write the inconsistency as a function of correlations and parameters (k and b_2). Therefore, we only need to derive k and b_2 as a function of correlations and then plug into previous equations.

From Eq.32a, k is the regression coefficient of M_U when we regress M_O on M_U and X. Similarly, from Eq.32c, b_2 is the regression coefficient of M_U when we regress Y on M_O, M_U and

X. Therefore, we can write the results in terms of correlations:

$$\begin{aligned}
\tilde{a}_1 &= a_1 + k \cdot \beta_1 \\
&= a_1 + \rho_{X,M_U} \cdot \frac{\rho_{M_O,M_U} - \rho_{X,M_O} \cdot \rho_{X,M_U}}{1 - \rho^2_{X,M_U}}
\end{aligned} \tag{43}$$

$$\begin{aligned}
\tilde{b}_1 &= b_1 + b_2 \cdot \beta_2 \\
&= b_1 + \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\cdot (\rho_{Y,M_U} + \rho_{Y,M_O} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} + \rho_{X,Y} \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_O} - \rho_{Y,M_U} \cdot \rho^2_{X,M_O} \\
&- \rho_{Y,M_O} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho_{X,M_U}) \cdot \frac{\rho_{M_O,M_U} - \rho_{X,M_U} \cdot \rho_{X,M_O}}{1 - \rho^2_{X,M_O}}
\end{aligned} \tag{44}$$

$$\begin{aligned}
\tilde{c} &= c + b_2 \cdot \beta_3 \\
&= c + \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\cdot (\rho_{Y,M_U} + \rho_{Y,M_O} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} + \rho_{X,Y} \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_O} - \rho_{Y,M_U} \cdot \rho^2_{X,M_O} \\
&- \rho_{Y,M_O} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho_{X,M_U}) \cdot \frac{\rho_{X,M_U} - \rho_{M_O,M_U} \cdot \rho_{X,M_O}}{1 - \rho^2_{X,M_O}}
\end{aligned} \tag{45}$$

Note: b_2 is derived based on formula of regression coefficients as follows (I derived both of these through the well-known $(X'X)^{-1}X'Y$ and some linear algebra).

When we regress Y on $X1$, $X2$ and $X3$, the coefficient of $X1$ is (in terms of correlations):

$$\begin{aligned}
&\frac{1}{1 + 2 \cdot \rho_{X1,X2} \cdot \rho_{X2,X3} \cdot \rho_{X1,X3} - \rho^2_{X1,X2} - \rho^2_{X2,X3} - \rho^2_{X1,X3}} \\
&\cdot (\rho_{Y,X1} + \rho_{Y,X2} \cdot \rho_{X1,X3} \cdot \rho_{X2,X3} + \rho_{Y,X3} \cdot \rho_{X1,X2} \cdot \rho_{X2,X3} - \rho_{Y,X1} \cdot \rho^2_{X2,X3} \\
&- \rho_{Y,X2} \cdot \rho_{X1,X2} - \rho_{Y,X3} \cdot \rho_{X1,X3})
\end{aligned} \tag{46}$$

Percent of inconsistency

In this section, we will derive the percent of inconsistency as a function of correlations.

From Eq.32 we know a_1, b_1 and c are also functions of correlations, which can be shown as follows.

$$\begin{aligned}
a_1 &= \frac{\rho_{X,M_O} - \rho_{M_O,M_U} \cdot \rho_{X,M_U}}{1 - \rho^2_{X,M_U}} \\
b_1 &= \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\quad \cdot (\rho_{Y,M_O} + \rho_{Y,M_U} \cdot \rho_{X,M_O} \cdot \rho_{X,M_U} + \rho_{X,Y} \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} - \rho_{Y,M_O} \cdot \rho^2_{X,M_U} \\
&\quad - \rho_{Y,M_U} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho_{X,M_O}) \\
c &= \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\quad \cdot (\rho_{X,Y} + \rho_{Y,M_U} \cdot \rho_{X,M_O} \cdot \rho_{M_O,M_U} + \rho_{Y,M_O} \cdot \rho_{X,M_U} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho^2_{M_O,M_U} \\
&\quad - \rho_{Y,M_U} \cdot \rho_{X,M_U} - \rho_{Y,M_O} \cdot \rho_{X,M_O})
\end{aligned}$$

Together with the previous section, we can get percent of inconsistency as a function of correlations.

$$\begin{aligned}
\frac{\tilde{a}_1 - a_1}{\tilde{a}_1} &= \frac{\rho_{X,M_U} \cdot (\rho_{M_O,M_U} - \rho_{X,M_O} \cdot \rho_{X,M_U})}{\rho_{X,M_O} \cdot (1 - \rho^2_{X,M_U})} \\
\frac{\tilde{b}_1 - b_1}{\tilde{b}_1} &= \frac{\rho_{M_O,M_U} - \rho_{X,M_O} \cdot \rho_{X,M_U}}{\rho_{Y,M_O} - \rho_{X,M_O} \cdot \rho_{X,Y}} \cdot \\
&\quad \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\quad \cdot (\rho_{Y,M_U} + \rho_{Y,M_O} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} + \rho_{X,Y} \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_O} - \rho_{Y,M_U} \cdot \rho^2_{X,M_O} \\
&\quad - \rho_{Y,M_O} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho_{X,M_U}) \\
\frac{\tilde{c} - c}{\tilde{c}} &= \frac{\rho_{X,M_U} - \rho_{M_O,M_U} \cdot \rho_{X,M_O}}{\rho_{X,Y} - \rho_{X,M_O} \cdot \rho_{Y,M_O}} \cdot \\
&\quad \frac{1}{1 + 2 \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} - \rho^2_{M_O,M_U} - \rho^2_{X,M_U} - \rho^2_{X,M_O}} \\
&\quad \cdot (\rho_{Y,M_U} + \rho_{Y,M_O} \cdot \rho_{X,M_U} \cdot \rho_{X,M_O} + \rho_{X,Y} \cdot \rho_{M_O,M_U} \cdot \rho_{X,M_O} - \rho_{Y,M_U} \cdot \rho^2_{X,M_O} \\
&\quad - \rho_{Y,M_O} \cdot \rho_{M_O,M_U} - \rho_{X,Y} \cdot \rho_{X,M_U})
\end{aligned}$$

Derive error variation for standardization

In order to have all variables standardized, we need to constrain the error variations. From Eq.32:

$$var(M_O) = var(k \cdot M_U) + var(a_1 \cdot X) + 2 \cdot cov(k \cdot M_U, a_1 \cdot X) + var(\epsilon_{M_O})$$

$$var(M_U) = var(a_2 \cdot X) + var(\epsilon_{M_U})$$

$$\begin{aligned} var(Y) = {} & var(b_1 \cdot M_O) + var(b_2 \cdot M_U) + var(c \cdot X) + \\ & 2 \cdot cov(b_1 \cdot M_O, b_2 \cdot M_U) + 2 \cdot cov(b_1 \cdot M_O, c \cdot X) + 2 \cdot cov(b_2 \cdot M_U, c \cdot X) + var(\epsilon_Y) \end{aligned}$$

$$var(\epsilon_{M_O}) = 1 - k^2 - a_1^2 - 2ka_1 \cdot \rho_{X,M_U} = 1 - k^2 - a_1^2 - 2ka_1a_2 \tag{49a}$$

$$var(\epsilon_{M_U}) = 1 - a_2^2 \tag{49b}$$

$$\begin{aligned} var(\epsilon_Y) &= 1 - b_1^2 - b_2^2 - c^2 - 2b_1b_2 \cdot \rho_{M_O,M_U} - 2b_1c \cdot \rho_{X,M_O} - 2b_2c \cdot \rho_{X,M_U} \\ &= 1 - b_1^2 - b_2^2 - c^2 - 2b_1b_2 \cdot (k + a_1a_2) - 2b_1c \cdot (ka_2 + a_1) - 2a_2b_2c \end{aligned} \tag{49c}$$

Note: From Eq.49, we can tell that we have following constraints on parameters:

$$1 - k^2 - a_1^2 - 2ka_1a_2 > 0 \tag{50a}$$

$$1 - a_2^2 > 0 \tag{50b}$$

$$1 - b_1^2 - b_2^2 - c^2 - 2b_1b_2 \cdot (k + a_1a_2) - 2b_1c \cdot (ka_2 + a_1) - 2a_2b_2c > 0 \tag{50c}$$

Discussion about the directions of inconsistency

In this section, we discuss the directions of inconsistency based on previous derivations.

with respect to a_1

From Eq.41a, we can tell $\tilde{a}_1 - a_1 > 0$ as long as k and a_2 have the same direction. That is, we overestimate a_1 when k and a_2 are both positive or both negative. If $k > 0, a_2 < 0$ or $k < 0, a_2 > 0$ we underestimate a_1.

with respect to b_1

From Eq.41b, we can show that the sign of $\tilde{b}_1 - b_1$ depends on whether k and b_2 have the same direction.

First, $1 - a^2 > 0$ based on Eq.50b.

Second, $1 - (k \cdot a_2 + a_1)^2 > 0$ because:

$$(k \cdot a_2 + a_1)^2 = k^2 \cdot a_2^2 + a_1^2 + 2ka_1a_2 < k^2 + a_1^2 + 2ka_1a_2 < 1$$

where the last inequality is based on Eq.50a.

Therefore, we overestimate b_1 when k and b_2 are both positive or both negative. If $k > 0, b_2 < 0$ or $k < 0, b_2 > 0$ we underestimate b_1.

with respect to c

From Eq.41c, we can tell that sign of $\tilde{c} - c$ depends on whether b_2 and $a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)$ have the same direction. This is because we have already shown that $1 - (k \cdot a_2 + a_1)^2 > 0$ when we discuss inconsistency for b_1.

We overestimate c when b_2 and $a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)$ are both positive or both negative. If one positive and one negative then we underestimate c.

Interestingly, from Eq.34 we can tell that:

$$a_2 - (k + a_1 \cdot a_2) = \rho_{X,M_U} - \rho_{M_O,M_U} \cdot \rho_{X,M_O}$$

Parameter framework: how inconsistency changes with parameters

In this section, we work out partial derivatives of $\tilde{a}_1$, $\tilde{b}_1$, $\tilde{c}$ with respect to parameters related to M_U: a_2, b_2 and k.

with respect to k

From Eq.41a, Eq.41b, Eq.41c

$$\frac{\partial \tilde{a}_1}{\partial k} = a_2 \tag{51a}$$

$$\frac{\partial \tilde{b}_1}{\partial k} = \frac{(1 - a_2^2)b_2 \cdot [a_2^2 k^2 + (1 - a_1^2)]}{[1 - (a_1 + a_2 k)^2]^2} \tag{51b}$$

$$\frac{\partial \tilde{c}}{\partial k} = \frac{b_2(1 - a_2^2)\{-2a_2 k + a_1[(a_1 + a_2 k)^2 - 1]\}}{(-1 + a_1^2 + 2a_1 a_2 k + a_2^2 k^2)^2} \tag{51c}$$

When all parameters are positive and follow the constraints given by Eq.50, it can be easily shown that $\frac{\partial \tilde{a}_1}{\partial k}$ and $\frac{\partial \tilde{b}_1}{\partial k}$ are always positive.

For $\frac{\partial \tilde{c}}{\partial k}$, we can show that it is always negative by following:

$$(a_1 + a_2 k)^2 = a_1^2 + a_2^2 k^2 + 2k a_1 a_2 < a_1^2 + k^2 + 2k a_1 a_2 < 1 \tag{52}$$

where the last inequality is based on Eq.50a.

with respect to a_2

From Eq.41a, Eq.41b, Eq.41c

$$\frac{\partial \tilde{a}_1}{\partial a_2} = k \tag{53a}$$

$$\frac{\partial \tilde{b}_1}{\partial a_2} = \frac{2b_2 k[a_1 k + a_1 a_2^2 k + a_2(-1 + a_1^2 + k^2)]}{[1 - (a_1 + a_2 k)^2]^2} \tag{53b}$$

$$\frac{\partial \tilde{c}}{\partial a_2} = \frac{b_2\{a_1^4 + 2a_1^3 a_2 k - 2a_1 a_2 k(1 + k^2) + (1 - k^2)(1 + a_2^2 k^2) + a_1^2[(a_2^2 - 1)k^2 - 2]\}}{[1 - (a_1 + a_2 k)^2]^2} \tag{53c}$$

When all parameters are positive, $\frac{\partial \tilde{a}_1}{\partial a_2}$ is positive.

When all parameters are positive, $\frac{\partial \tilde{b}_1}{\partial a_2}$ and $\frac{\partial \tilde{c}}{\partial a_2}$ can be either positive or negative, depending on the magnitudes of a_1 and k. Discussions are as follows.

Fig..8 shows how the sign of $\frac{\partial \tilde{b}_1}{\partial a_2}$ changes when a_2, a_1 and k take on different values. The dashed line represents the constraint given by Eq.50a. The constraint says that we can only take on values below the dashed line. The solid line, on the other hand, shows where $\frac{\partial \tilde{b}_1}{\partial a_2}$ is equivalent to

0. Importantly, this partial derivative is positive when a_1 and k take on values above the solid line and negative below the solid line. And these two lines become closer to each other as a_2 becomes larger, which are shown from the left figure to the right one.

Similarly, Fig..9 shows how the sign of $\frac{\partial \tilde{c}}{\partial a_2}$ changes when a_2, a_1 and k take on different values. Again, the dashed line represents the constraint given by Eq.50a. And we can only take on values below the dashed line. The solid line shows where $\frac{\partial \tilde{c}}{\partial a_2}$ is equivalent to 0. Above this solid line this partial derivative is negative and below the solid line it is positive. And the three figures from left to right show what happens as a_2 becomes larger.

Therefore, we can summarize as follows:

When a_1 and k are relatively small (in the left-lower corner), $\frac{\partial \tilde{b}_1}{\partial a_2}$ can be positive first then quickly become negative as a_2 becomes larger. Instead, $\frac{\partial \tilde{b}_1}{\partial a_2}$ is always positive.

When a_1 and k are relatively large (in the center part), $\frac{\partial \tilde{b}_1}{\partial a_2}$ can be always positive. For $\frac{\partial \tilde{b}_1}{\partial a_2}$, the sign can be negative or first positive then negative.

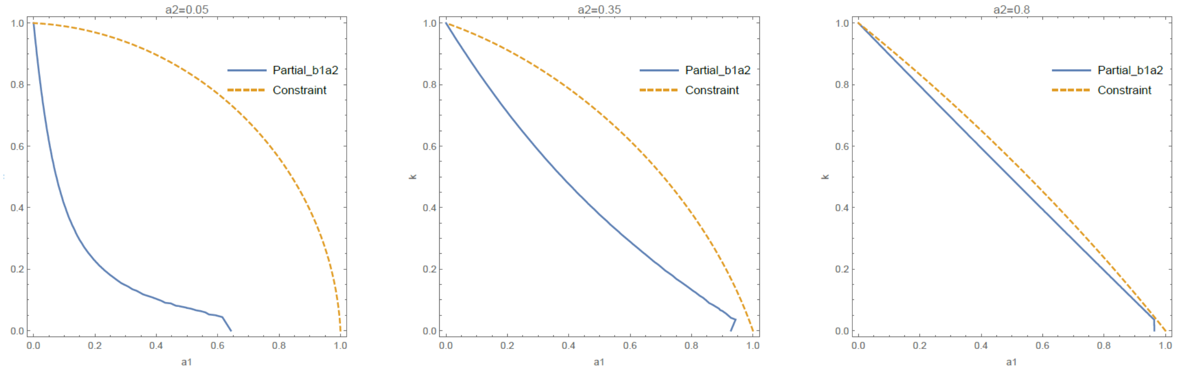


Figure .8: Sign of $\frac{\partial \tilde{b}_1}{\partial a_2}$

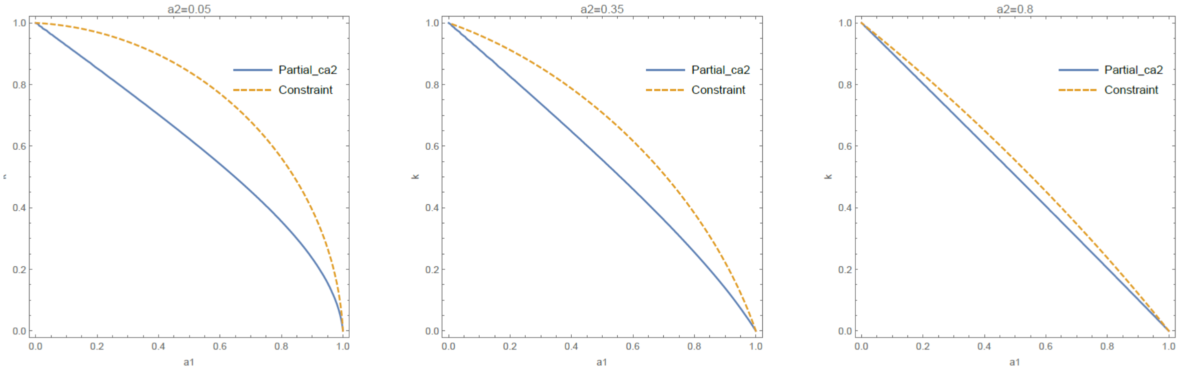


Figure .9: Sign of $\frac{\partial \tilde{c}}{\partial a_2}$

with respect to b_2

From Eq.41a, Eq.41b, Eq.41c

$$\frac{\partial \tilde{a}_1}{\partial b_2} = 0 \tag{54a}$$

$$\frac{\partial \tilde{b}_1}{\partial b_2} = k \cdot \frac{1 - a_2^2}{1 - (k \cdot a_2 + a_1)^2} \tag{54b}$$

$$\frac{\partial \tilde{c}}{\partial b_2} = \frac{a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)}{1 - (k \cdot a_2 + a_1)^2} \tag{54c}$$

When all parameters are positive (and other restrictions required by standardization), it can be shown that $\frac{\partial \tilde{b}_1}{\partial b_2}$ is always positive by using Eq.52.

$\frac{\partial \tilde{c}}{\partial b_2}$ can be either positive or negative, depending on the magnitudes of a_1, a_2 and k.

Fig..10 shows how the sign of $\frac{\partial \tilde{b}_1}{\partial a_2}$ changes when a_2, a_1 and k take on different values. The dashed line represents the constraint given by Eq.50a. The constraint says that we can only take on values below the dashed line. The solid line, on the other hand, shows where $\frac{\partial \tilde{c}}{\partial b_2}$ is equivalent to 0. This partial derivative is positive when a_1 and k take on values below the solid line and negative above the solid line. The three figures from left to right show what happens when a_2 turns larger.

Therefore, when a_1 and k are relatively small (in the left-lower corner) and a_2 is also very small, $\frac{\partial \tilde{c}}{\partial a_2}$ is positive. When a_1 and k are relatively large (compared to a_2)(in the center part, between the two lines), $\frac{\partial \tilde{c}}{\partial a_2}$ is negative. Importantly, because we are talking about the partial derivatives with regards to c and c does not appear in this derivative, the slope always keeps the same no matter what value c takes.

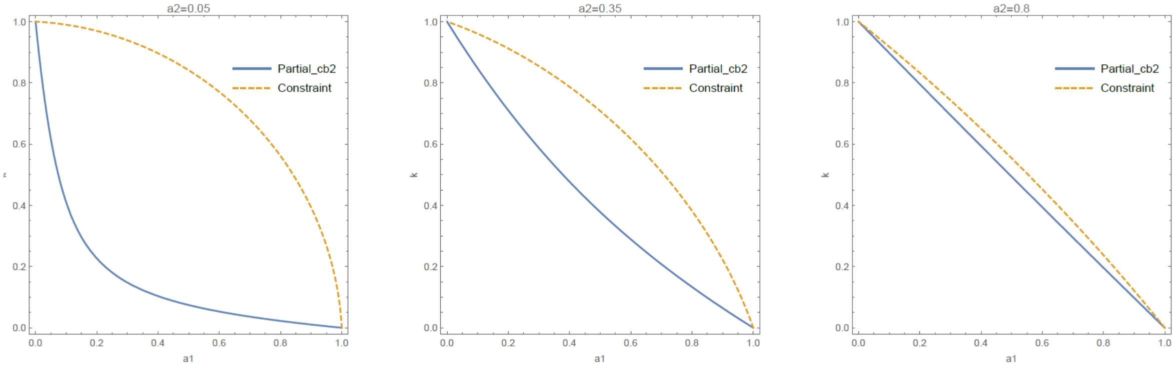


Figure .10: Sign of $\frac{\partial \tilde{c}}{\partial b_2}$

Decompose total inconsistency into two missing pathways when omitting M_U

First we write the following equation based on the results in Eq. 41.

$$[\tilde{a}_1 \cdot \tilde{b}_1 - a_1 \cdot b_1] + [\tilde{c} - c] = a_2 \cdot k \cdot b_1 + b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) + b_2 \cdot \beta_3 \tag{55}$$

Then following derivation shows that the last two components $b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) + b_2 \cdot \beta_3$ is equivalent to $a_2 \cdot b_2$.

$$\begin{aligned}
&b_2 \cdot \beta_2 \cdot (a_1 + k \cdot a_2) + b_2 \cdot \beta_3 && \text{(56a)}\\
&= b_2 \cdot k \cdot \frac{1 - a_2^2}{1 - (k \cdot a_2 + a_1)^2} \cdot (a_1 + k \cdot a_2) + b_2 \cdot \frac{a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)}{1 - (k \cdot a_2 + a_1)^2}\\
&= \frac{b_2 \cdot k \cdot (1 - a_2^2) \cdot (a_1 + k \cdot a_2) + b_2 \cdot [a_2 - (k + a_1 \cdot a_2) \cdot (k \cdot a_2 + a_1)]}{1 - (k \cdot a_2 + a_1)^2}\\
&= \frac{(k \cdot a_2 + a_1) \cdot b_2 \cdot [k \cdot (1 - a_2^2) - (k + a_1 \cdot a_2)] + b_2 \cdot a_2}{1 - (k \cdot a_2 + a_1)^2}\\
&= \frac{(k \cdot a_2 + a_1) \cdot b_2 \cdot [k - k \cdot a_2^2 - k - a_1 \cdot a_2] + b_2 \cdot a_2}{1 - (k \cdot a_2 + a_1)^2}\\
&= \frac{-(k \cdot a_2 + a_1) \cdot b_2 \cdot a_2 \cdot (k \cdot a_2 + a_1) + b_2 \cdot a_2}{1 - (k \cdot a_2 + a_1)^2}\\
&= a_2 \cdot b_2 && \text{(56b)}
\end{aligned}$$

REFERENCES

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of labor Economics*, *25*(1), 95–135.

Adnot, M., Dee, T., Katz, V., & Wyckoff, J. (2017). Teacher Turnover, Teacher Quality, and Student Achievement in DCPS. *Educational Evaluation and Policy Analysis*, *39*(1), 54-76.

An, W. (2018). Causal Inference with Networked Treatment Diffusion. *Sociological Methodology*, *48*(1), 152–181.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bekman, N. M., Cummins, K., & Brown, S. A. (2010). Affective and personality risk and cognitive mediators of initial adolescent alcohol use. *Journal of studies on alcohol and drugs*, *71*(4), 570–580.

Blatchford, P., & Martin, C. (1998). The Effects of Class Size on Classroom Processes: 'It's a Bit Like a Treadmill – Working Hard and Getting Nowhere Fast!'. *British Journal of Educational Studies*, *46*(2), 118–137.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842.

Bloom, H. S., & Spybrook, J. (2017). Assessing the Precision of Multisite Trials for Estimating the Parameters of a Cross-Site Population Distribution of Program Effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902.

Bulterman-Bos, J. A. (2008). Will a Clinical Approach Make Education Research More Relevant for Practice? *Educational Researcher; Washington*, *37*(7), 412–420.

Carrell, S. E., Fullerton, R. L., & West, J. E. (2009). Does Your Cohort Matter? Measuring Peer Effects in College Achievement. *Journal of Labor Economics*, *27*(3), 439–464.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. *Working Paper*(17699).

Cinelli, C., & Hazlett, C. (2018). Making Sense of Sensitivity: Extending Omitted Variable Bias. *Working Paper*, 40.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, New Jersey: Erlbaum Associates.

Cook, T. D. (2002, September). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Dee, T. S., & Wyckoff, J. (2015). Incentives, Selection, and Teacher Performance: Evidence from IMPACT: Incentives, Selection, and Teacher Performance. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Dieterle, S., Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). How do Principals Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value Added: How do Principals Assign Students to Teachers? *Journal of Policy Analysis and Management*, *34*(1), 32–58.

Frank, K. A. (1998). Chapter 5: Quantitative Methods for Studying Social Context in Multilevels and Through Interpersonal Relations. *Review of Research in Education*, *23*(1), 171–216.

Frank, K. A. (2000). Impact of a Confounding Variable on a Regression Coefficient. *Sociological Methods & Research*, *29*(2), 147–194.

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., Saw, G. K., & Xu, R. (2016). Book review of "Causality in a Social World" by Guanglei Hong. *Observational Studies*, 4.

Fritz, M. S., Kenny, D. A., & MacKinnon, D. P. (2016). The Combined Effects of Measurement Error and Omitting Confounders in the Single-Mediator Model. *Multivariate Behavioral Research*, *51*(5), 681–697.

Fritz, M. S., & MacKinnon, D. P. (2007). Required Sample Size to Detect the Mediated Effect. *Psychological Science*, *18*(3), 233–239.

Goldhaber, D., & Hansen, M. (2010). Assessing the Potential of Using Value-Added Estimates of Teacher Job Performance for Making Tenure Decisions. Working Paper 31. *National Center for Analysis of Longitudinal Data in Education Research*.

Gordon, R. J., Kane, T. J., & Staiger, D. (2006). *Identifying effective teachers using performance on the job*. Brookings Institution Washington, DC.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2015). Can value-added measures of teacher performance be trusted? *Education Finance and Policy*.

Hanushek, E. A. (1999). Some Findings From an Independent Investigation of the Tennessee STAR Experiment and From Other Investigations of Class Size Effects. *Educational Evaluation and Policy Analysis*, *21*(2), 143–163.

Hanushek, E. A. (2014). Boosting teacher effectiveness. *What lies ahead for America's children and their schools*, 23–35.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

Harfitt, G. J. (2013). Why 'small' can be better: an exploration of the relationships between class size and pedagogical practices. *Research Papers in Education*, *28*(3), 330–345.

Harris, D. N. (2009). Teacher value-added: Don't end the search before it starts. *Journal of Policy Analysis and Management*, *28*(4), 693–699.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*. New York: The Guilford Press.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794–831.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hong, G. (2015). *Causality in a Social World: Moderation, Meditation and Spill-over*. Chichester, UK: John Wiley & Sons, Ltd.

Hong, G., Qin, X., & Yang, F. (2018). Weighting-Based Sensitivity Analysis in Causal Mediation Studies. *Journal of Educational and Behavioral Statistics*, *43*(1), 32–56.

Hoxby, C. M., & Weingarth, G. (2005). TAKING RACE OUT OF THE EQUATION: SCHOOL REASSIGNMENT AND THE STRUCTURE OF PEER EFFECTS. *Working paper*, 47.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*(4), 309–334.

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, *25*(1), 51–71.

Imai, K., & Yamamoto, T. (2013). Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments. *Political Analysis*, *21*(2), 141–171.

Imbens, G. W. (2003). Sensitivity to Exogeneity Assumptions in Program Evaluation. *American*

Economic Review, *93*(2), 126–132.

Judd, C. M., & Kenny, D. A. (1981). Process Analysis: Estimating Mediation in Treatment Evaluations. *Evaluation Review*, *5*(5), 602–619.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Kim, C. M., Frank, K. A., & Spillane, J. P. (2018). Relationships among teachers' formal and informal positions and their incoming student composition. *Teachers College Record*, *120*(3), n3.

Konstantopoulos, S. (2011). How Consistent Are Class Size Effects? *Evaluation Review*, *35*(1), 71–92.

Levin, J. (2002). For whom the reductions count: A quantile regression analysis of class size and peer effects on scholastic achievement. In *Economic applications of quantile regression* (p. 221-246). Springer.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of educational and behavioral statistics*, *27*(3), 255–270.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement*, *44*(1), 47–67.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York: Lawrence Erlbaum Associates. (OCLC: ocn122701406)

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention science*, *1*(4), 173–181.

Maroulis, S. (2016). Interpreting school choice treatment effects: Results and implications from computational experiments. *Journal of Artificial Societies and Social Simulation*, *19*(1), 7.

Maroulis, S., Guimera, R., Petry, H., Stringer, M. J., Gomez, L. M., Amaral, L. A. N., & Wilensky, U. (2010). Complex Systems View of Educational Policy Research. *Science*, *330*(6000), 38–39.

Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, *12*(1), 23–44.

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in Cross-Sectional Analyses of

Longitudinal Mediation: Partial and Complete Mediation Under an Autoregressive Model. *Multivariate Behavioral Research*, *46*(5), 816–841.

Mitchell, M. A., & Maxwell, S. E. (2013). A Comparison of the Cross-Sectional and Sequential Designs when Assessing Longitudinal Mediation. *Multivariate Behavioral Research*, *48*(3), 301–339.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). The Effects of Small Classes on Academic Achievement: The Results of the Tennessee Class Size Experiment. *American Educational Research Journal*, *37*(1), 123–151.

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

Paufler, N. A., & Amrein-Beardsley, A. (2014). The random assignment of students into elementary classrooms: Implications for value-added analyses and interpretations. *American Educational Research Journal*, *51*(2), 328–362.

Pedder, D. (2006). Are small classes better? Understanding relationships between class size, classroom processes and pupils' learning. *Oxford Review of Education*, *32*(2), 213–234.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891.

Raudenbush, S. W. (2015). Value added: A case study in the mismatch between education research and policy. *Educational Researcher*, *44*(2), 138–141.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer.

Rosenberg, M. (1968). *The logic of survey analysis*. New York: Basic.

Rothstein, J. (2009, October). Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables. *Education Finance and Policy*, *4*(4), 537–571.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1986). Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *Journal of the American Statistical Association*, *81*(396), 961.

Rubin, D. B. (1990). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, *25*(3), 279–292.

Schanzenbach, D. W. (2007). What Have Researchers Learned from Project STAR? *Brookings Papers on Education Policy*(9), 205–228.

Singh, R., Chen, F., & Wegener, D. T. (2014). The Similarity-Attraction Link: Sequential Versus Parallel Multiple-Mediator Models Involving Inferred Attraction, Respect, and Positive Affect. *Basic and Applied Social Psychology*, *36*(4), 281–298.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, *13*, 290–312.

Spybrook, J., & Raudenbush, S. W. (2009). An Examination of the Precision and Technical Accuracy of the First Wave of Group-Randomized Trials Funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis*, *31*(3), 298–318.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267.

Sullivan, G. M. (2011). Getting Off the "Gold Standard": Randomized Controlled Trials and Education Research. *Journal of Graduate Medical Education*, *3*(3), 285–289.

Tal-Or, N., Cohen, J., Tsfati, Y., & Gunther, A. C. (2010). Testing Causal Direction in the Influence of Presumed Media Influence. *Communication Research*, *37*(6), 801–824.

VanderWeele, T. J. (2010). Bias Formulas for Sensitivity Analysis for Direct and Indirect Effects:. *Epidemiology*, *21*(4), 540–551.

VanderWeele, T. J. (2015). *Explanation in causal inference: methods for mediation and interaction*. New York: Oxford University Press.

VanderWeele, T. J., & Arah, O. A. (2011). Bias Formulas for Sensitivity Analysis of Unmeasured Confounding for General Outcomes, Treatments, and Confounders:. *Epidemiology*, *22*(1), 42–52.

VanderWeele, T. J., & Chiba, Y. (2014). Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiology, Biostatistics and Public Health*(ONLINE FIRST).

Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, *5*(2), 134-150.

Winters, M. A., & Cowen, J. M. (2013a). Who Would Stay, Who Would Be Dismissed? An Empirical Consideration of Value-Added Teacher Retention Policies. *Educational Researcher*, *42*(6), 330–337.

Winters, M. A., & Cowen, J. M. (2013b). Would a Value-Added System of Retention Improve the Distribution of Teacher Quality? A Simulation of Alternative Policies: Value Added Teacher Retention Policies. *Journal of Policy Analysis and Management*, *32*(3), 634–654.

Wooldridge, J. M. (2009). *Introductory econometrics: a modern approach* (4th ed ed.). Mason, OH: South Western, Cengage Learning.