ABSTRACT

Title of Dissertation: ESTIMATING THE LONGITUDINAL COMPLIER AVERAGE CAUSAL EFFECT USING THE LATENT GROWTH MODEL: A SIMULATION STUDY

Huili Liu, Doctor of Philosophy, 2018

Dissertation directed by: Professor, Gregory Hancock
Professor, Laura Stapleton
Measurement, Statistics, and Evaluation
Department of Human Development and Quantitative Methodology

When noncompliance happens to longitudinal experiments, the randomness for drawing causal inferences is contaminated. In such cases, the longitudinal Complier Average Causal Effect (CACE) is often estimated. The Latent Growth Model (LGM) is very useful in estimating longitudinal trajectories and can be easily adapted for estimating longitudinal CACE.

Two popular CACE approaches, the Standard IV approach and the Mixture Model Based (MMB) approach, are both readily applicable to the LGM framework. The Standard IV approach is simple in modeling and has low computational burden, but it is also criticized for ignoring distributions of subgroups and leading to biased estimations. The MMB approach is capable of not only estimating the CACE but also answering research questions regarding distributions of subpopulations, but this

method may yield unstable results under unfavorable conditions, especially when the estimation model is complicated.

Previous studies laid out a theoretical background for applying LGMs to longitudinal CACE estimation using both approaches. However, 1) very little was known regarding the factors that might influence the longitudinal CACE estimation, 2) the three compliance classes scenario was not thoroughly investigated, and 3) it was still unclear about how and to what extent the Standard IV approach would perform better or worse than the MMB approach in the longitudinal CACE estimation.

The present study used an intensive simulation design to investigate the performance of the Standard IV and the MMB approaches while manipulating six factors that were related to most experimental designs: sample size, compliance composition, effect size, reliability of measurements, mean distances, and noncomplier-complier Level 2 covariance ratio. Their performance was evaluated on four criteria, estimation success rate, estimation bias, power, and type I error rate. With the analysis result, suggestions regarding experiment designs were provided for researchers and practitioners.

ESTIMATING THE LONGITUDINAL COMPLIER AVERAGE CAUSAL EFFECT USING THE LATENT GROWTH MODEL: A SIMULATION STUDY

by

Huili Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Gregory Hancock, Co-Chair
Professor Laura Stapleton, Co-Chair
Professor Jeffrey Harring
Professor Ji Seung Yang
Professor Hedwig Teglasi, Dean's Representative

Dedication

This dissertation is dedicated to my family and my husband, Mike.

Acknowledgements

I am extremely grateful for my two advisors, Dr. Gregory Hancock and Dr. Laura Stapleton. They have always been supportive and helpful when I ran into different obstacles and struggled for assistance. Because I was working and writing my dissertation at the same time, they not only showed understanding but also gave extra stipulations to keep me on track. Whenever I had questions and needed guidance, they were always ready to provide great insights. I am also extremely appreciative for their support throughout my Ph.D. journey. They have provided great opportunities and assistance for me to establish my professional development. I am deeply indebted to them for their help.

I want to thank Dr. Jeffrey Harring, Dr. Ji Seung Yang, and Dr. Hedwig Teglasi for reviewing my dissertation. They reviewed the preliminary text of my dissertation and provided valuable suggestions on improvement. With the help of their great perceptions, I was able to have a clearer view of the big picture and pay greater attentionaa to details.

I wish to thank all my colleagues and friends at the EDMS program for providing emotional encouragement and having professional discussions with me. Through discussing the problems I have encountered with them, I was able to sort out a large number of issues and continue my momentum in my research. I am extremely thankful for having such a great group of resourceful and helpful friends

Throughout this process my family and my friends have been a tremendous source of support and assistance. I wish to thank my greatest husband, Mike, not only for being the most wonderful emotional supporter in the world but also for his

guidance on coding and parallel computing. I am also grateful for all my family members, both my side and my husband's side, for being understanding and supportive. I will also address my deepest gratitude to my friends, Dandan, Xi, Fei, Xin, and Ying. With you all, I shared my sorrow and happiness throughout this journey, and I am awfully lucky to have you.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Educational research is witnessing an upsurge of announcements declaring the most innovative and effective methods in improving education systems, enhancing student performance, boosting social development, and better preparing the younger generation for modern challenges. However, public resources are not ample enough to test every new proposal, so it is extremely critical for policymakers to scrutinize all study results and identify the ones that are meticulously conducted with proper scientific methodologies.

1.1.Randomized Experiments

Causal inference is a highly valued scientific approach in many research areas. Economists need to define and identify causal parameters to address policy issues (Heckman, 2008); biologists strive to disentangle complex patterns to search for the ultimate causal processes that can actually revolutionize agricultural or medical research (Shipley, 2002); for social study researchers, pinpointing causes can be a cure for problems that blight our social development. On one hand, it is a human inclination to think in causal terms; on the other, only with laborious arguments searching for the causal chain can one rule out confounding alternatives and demystify and highlight the veiled truth (Einhorn & Hogarth, 1986). Among various methods for drawing causal inferences, the randomized experimental design was considered as a "gold standard" (Shadish, Cook, & Campbell, 2002, p. 13).

A randomized experiment is defined as an experiment in which units are randomly assigned to either the treatment group or the control group (Shadish et al.,

2002). As Shadish et al. (2002) argued, randomized experimental designs were widely acclaimed in various areas because they promised control over confounding factors by creating two or more groups that were probabilistically equivalent on expectation even without physical isolation in laboratories. In this way, the existence and the magnitude of the treatment-control group difference can be credited to the treatment rather than pre-existed group differences before implementing an experiment.

Many educational researchers (e.g., Boruch, De Moya & Snyder, 2002; Mosteller & Boruch, 2002; Slavin, 2002) have shown strong support for the use of randomized experiments for evaluations of educational interventions and policies. Slavin (2002) argued that randomized experiments would be applicable for educational programs "in every subject and every grade level" (p. 17) and for a large variety of important topics such as "school-to-work transitions, special education, gifted children, dropout prevention, English language learners, race relations, drug abuse prevention, violence prevention" (p. 7). These studies may involve political decisions that are as expensive as millions of dollars, or more importantly, they could help students to build a more successful future.

Historically, educational studies did not really embrace this powerful approach. In his *Dewitt Wallace-Reader's Digest Distinguished Lecture*, Slavin (2002) lamented that while other academic communities had been "transformed" (p. 15) by the use of randomized experiments in the 20th century, educational research was "finally entering the 20th century" (p. 15) at the beginning of the 21st century. He pointed out that, using a 6-point scale—Strong, Promising, Marginal, Mixed, Weak,

or No Effect (Herman, 1999)—to evaluate the effectiveness of 2,665 Comprehensive School Reform (CSR) grants awarded between 1998 and 2002, only 20.8% reached the level of having Strong evidence of effectiveness, and 63.2% did not even reach the standards of having Marginal evidence of effectiveness. At the same time, state officials who reviewed the CSR grants proposals were not showing determination to tighten up the standards for scientific evidence.

Recognizing the urgency for educational research to have a scientific revolution, and the sluggish reaction among government officials and researchers, in 2002 the U.S. Congress established the Institute of Education Science (IES). One clear duty for the IES is to promote "scientifically valid research in education" (H.R. 3081, 2002, p. 6). In their later report, the IES once again emphasized their resolution of promoting the use of scientific designs, "especially randomized designs" (Whitehurst, 2008, p. 3).

1.2.Noncompliance in Randomized Experiments

There are multiple reasons that randomized experiments are formidable for researchers. Sometimes, randomized experiments are precluded because of legal, ethical (e.g., it is illegal and unethical to force participants to smoke for a smoking study), or logistical (e.g., it is hard to keep track of each individual's group membership in a large scale multiyear project that involves a lot of unexpected changes) reasons (Shadish, Clark, & Steiner, 2008). Other times, researchers may be reluctant to adhere to stringent methodological principles because of practical challenges, such as higher research cost and limited resources (Hsieh et al., 2005). In addition, even with a successful randomization, another issue can make the estimation

of treatment effects problematic—noncompliance of assignment (Angrist, Imbens, & Rubin, 1996; Jo & Muthén, 2001; Little & Yau, 1998).

Noncompliance of assignment arises when study participants fail to follow a randomized group assignment predetermined by researchers during the implementation of an experiment. One example of noncompliance is the JOBS II intervention study (Vinokur, Price, & Schul, 1995). This study was designed to improve mental health and facilitate reemployment for the recently unemployed through an intervention. At baseline, eligible participants were selected and randomly assigned to the control and treatment groups. The treatment group was provided with five job search training seminars, and the control group was only provided with a self-guided job-searching booklet. Although the researchers managed to achieve randomness before the study, only 54% of the treatment group members showed up in the seminars. In other words, only 54% of the treatment group members complied with their assignments, and the other 46% self-selected out of the intervention. In this study, control group members had no access to the seminar, so all control group members complied with their assignments. If the seminars were completely accessible to all individuals, it is possible to observe some control group members taking the treatment and therefore not complying with their assignments.

With non-compliance, 1) if one just simply estimates the treatment effect using the mean difference between the treatment and control groups, the result is an estimation of the effect of the treatment *assignment* not the effect of the intervention, and 2) if one instead uses the mean difference between individuals who actually take

the treatment and those who do not, the treatment effect estimation will be confounded with selection bias (Little & Yau, 1998).

There is a third choice when noncompliance happens: the Complier-Average Causal Effect (CACE) approach (Imbens & Rubin, 1997a, 1997b). The CACE approach is a special version of the instrumental variable analysis developed by Nobel-prize-winning economist James Heckman (1978).

For instrumental variable analyses, the main independent variable is usually endogenous. In other words, the independent variable *cannot* be guaranteed to be entirely random by research designs; instead, it is contaminated by some internal factors related to the outcome variable over which researchers have no control. If there exists an exogenous instrumental variable that covaries with the independent variable and satisfies certain restrictions (Angrist et al., 1996), the instrumental variable can carve out the exogenous variation (i.e., the part decided by the instrumental variable) in the main independent variable and estimate its causal effect on the outcome variable with only this part (Murnane & Willett, 2010). The effect estimated with only the exogenous part is called the Local Average Treatment Effect (LATE) (Angrist et al., 1996). Exogenous here means that the variable is not in any way affected by participants in the experiment, and it should be determined by some external sources.

The more general instrumental variable approach utilizes a wide selection of instrumental variables. In his study of the impact of educational attainment on individual's civic engagement, Dee (2004) used a continuous variable, the distance between individual's high school and the nearest two-year college, as the instrumental

variable. Angrist, Bettinger, Bloom, King, and Kremer (2002) used a binary variable, an exogenous offer of a scholarship, as the instrumental variable to estimate the LATE of using a scholarship on educational attainment.

In the study of Angrist et al. (2002), low-income families (i.e., experiment subjects) in Columbia could win a scholarship (i.e., treatment) that helped to pay for tuition at private secondary schools by participating a lottery. Some scholarship recipients sent their children to a private school because of the financial aid, while other recipients did not do so; some nonrecipients could only enroll their children to a public school because of no financial aid, but other nonrecipients chose private schools nonetheless. If the research question was to investigate the effect of using the scholarship on educational attainment, the independent variable, using the scholarship, was not exogenous. However, the random assignment of the scholarship was exogenous and could thus serve as the instrumental variable to carve out the exogenous part of the independent variable. This is a special case of the LATE approach, and it is essentially estimating the average treatment effect for the compliers or the Complier-Average Causal Effect (CACE) (Imbens & Rubin, 1997a, 1997b). In other words, using the random assignment variable, participants who complied with the assignment (i.e., compliers) were separated from participants who did not comply (i.e., non-compliers), and the treatment effect of the scholarship was estimated using only the compliers. More technical details about the CACE approach can be found in Chapter 2.

Imbens (2014) described the estimation of the CACE as "an analysis in a second-best setting" (p. 20). When noncompliance nonetheless happens in a

randomized experiment, researchers are not able to answer the original research question with total credibility and precision. In this case, compliers are the only subpopulation that researchers have full confidence in identifying the average treatment effect for. With the estimated CACE, researchers are able to check if an intervention or a policy has an effect on the population who would actually receive the treatment.

Two main estimation approaches are widely used to estimate the CACE. The first approach, the standard instrumental variable (Standard IV) estimation approach was first introduced by Bloom (1984) and then fully developed by Angrist et al. (1996) and Angrist and Imbens (1995). This approach does not incorporate the distribution of the outcome variable, and it is relatively easy to calculate. The second approach, by using the mixture model based (MMB) estimation method (Imbens & Rubin, 1997b) or a Bayesian framework (Imbens & Rubin, 1997a), can estimate the marginal distribution of the outcome variable for the complier subpopulation. Therefore, researchers are able to investigate not only the average difference between the treatment and control groups but also the whole distribution of the outcome variable. The main difference between these two approaches is that the MMB approach imposes nonnegativity on the marginal distribution of the outcome variables while the Standard IV approach is distribution free. In other words, the Standard IV approach only makes an unrestricted estimate of group mean difference, and the nonnegative outcome distribution cannot be imposed (Imbens & Rubin, 1997b).

Although Imbens and Rubin (1997b) argued that the Standard IV method could have a more biased result when comparing to the MMB approach, both

approaches were widely used in practice (e.g., Frumento, Mealli, Pacini, & Rubin, 2012; Hirano, Imben, Rubin, & Zhou, 2000; Sussman & Hayward, 2010).

1.3.Longitudinal Experimental Designs

When conducting educational experiments, such as a psychological intervention or an implementation of a new educational policy, the treatment effect can be either short term or long term depending on the research purpose. When immediate results are the focus of a study, short-term experiments should be used. In other cases, some experiments might have long-term effects, some experiments might have different short-term and long-term effects, and some experiments might have an unclear timeline about the onset of the treatment effects (Farrington, Loeber, & Welsh, 2010). In these situations, outcome variables should be measured repeatedly to collect longitudinal data. This type of design is the longitudinal experimental design.

Longitudinal designs involve at least two data collection points on the same individual and therefore allow more investigation of individual change (Fitzmaurice, Laird, & Ware, 2011). Compared to cross-sectional studies where response variables are measured only once, longitudinal designs have certain advantages, such as providing information about the development of outcome variables, helping to understand the onset of treatment effects, and displaying within-subject changes (Farrington et al., 2010). Therefore, if an experimental study focuses on finding any one or more of the listed features above, a longitudinal experimental design should be implemented.

1.3.1. Latent growth models

When analyzing longitudinal data, there are multiple widely used methods, including some more traditional methods (e.g., analysis of variance [ANOVA], multivariate analysis of variance [MANOVA], analysis of covariance [ANCOVA], auto-regressive and cross-lagged multiple regression) and some newer methods emerging from the structural equation modeling (SEM) area, such as latent growth models [LGMs] (Hancock, Harring, & Lawrence, 2013). Despite their popularity among researchers, traditional longitudinal data-analysis techniques have significant drawbacks. Hancock et al. (2013) explained that traditional methods were "somewhat circumspect" (p. 172) with less ideal data structures and with hypotheses that focused on individual-level changes. LGMs, on the other hand, do not suffer from these limitations.

LGMs can also accommodate external variables to predict different latent intercepts and trajectories across different levels of the variables (Hancock et al., 2013), so the LGM technique can easily adapt to longitudinal experimental scenarios by introducing the treatment variable into the model as a predictor or by using multi-group LGM to estimate the slope difference in the treatment and the control groups. Muthén and Curran (1997) defined the treatment effect of a longitudinal experiment under the framework of LGM as the difference in the growth trajectories of the treatment and control groups when randomization and treatment implementation were both conducted at the baseline. Similar to most SEM techniques, with a correct model specification, LGM approach can also improve the precision of the estimated

treatment effect with its ability to handle measurement errors at each time point (Jo & Muthén, 2003).

1.3.2. Noncompliance in longitudinal experimental designs

When conducting a longitudinal experiment, there are several variations for researchers to adapt to their specific research designs and data structure. Gao, Brown, and Elliott (2014) summarized three key factors: a) randomization occurrence, b) treatment implementation, and c) subjects' compliance status assumption. In more detail, a) researchers can randomize their subjects to different treatment levels only once at the baseline (e.g., Vinokur et al., 1995) or more than once at different time points (e.g., Frangakis et al., 2004); b) treatment can be implemented once at the baseline (e.g., Vinokur et al., 1995) or multiple times over time (e.g., Frangakis et al., 2004); c) with more than one implementation of a treatment, researchers may assume subjects' compliance status to be time-invariant (e.g., Yau & Little, 2001) or time-varying (e.g., Lin, Have, & Elliott, 2009) across all implementations.

Given the current status quo of educational research, where randomization is difficult to implement, the present study only discussed baseline randomization. In addition, for most educational policy or intervention, one-time treatment application is fairly common (e.g., Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002; Harris & Goldrick-Rab, 2012; Vinokur et al., 1995), so only the baseline treatment implementation was considered. Lastly, because only one-time implementation is included, there was no need to consider if subjects' compliance statuses changed over time.

1.3.3. Latent growth model with noncompliance

When noncompliance occurs, with the assumption of baseline randomization and treatment implementation and the assumption of time-invariant compliance status, the CACE can be defined as the difference in the growth trajectories of compliers in treatment and control groups. Previous research (Jo & Muthén, 2003; Muthén, 2002) has explained how to apply the LGM technique to estimate the longitudinal CACE using the MMB estimation method.

For example, Jo and Muthén (2003) used the data from the Johns Hopkins Public School Preventive Intervention Study (JHU PIRC) (Ialongo et al., 1999) and estimated the longitudinal CACE with the MMB estimation method. This study was designed to investigate the respective effects of two interventions (i.e., classroom-centered [CC] intervention and family-school partnership [FSP] intervention), when comparing to the same control condition, on improving school children's academic achievement and reducing delinquencies. Jo and Muthén (2003) only focused on the comparison between the FSP intervention group and the control group.

The FSP intervention required parents to implement 66 take-home activities within 6 months while the control group was not asked to do any of these activities. Students were assessed at three time points: before the intervention (Month 0), right after the intervention (Month 6), and 12 months after the intervention (Month 18). Parents and students were first randomly assigned to the intervention and the control groups. Noncompliance occurred because only part of the parents completed all 66 activities. Therefore, for the intervention group, parents who completed at least 35 activities were classified "compliers" and the rest were classified as "non-compliers".

For the control group, none of the individuals had access to the intervention, so there was no "non-complier".

With the MMB based estimation of the CACE, they were able to estimate the unknown compliance status for each subject in the control group. Combining with the LGM framework, they were able to estimate the intervention-control group difference in trajectories among compliers. As a result, they found the intervention group had a significantly steeper trajectory than the control group. More technical details are included in Chapter 2.

1.4.Purpose and Research Questions

Previous studies have laid a theoretical background for applying LGM techniques to estimate longitudinal CACE using the MMB estimation method. However, very little is known about factors that might influence the accuracy and efficiency of the longitudinal CACE estimation. Jo (2002) investigated the influence of several factors (compliance rate, study design, outcome distributions, and covariate information) on statistical power for randomized interventions with noncompliance, but she only considered cross-sectional designs where the outcome variable was measured only once. It is relatively unclear about how these factors can affect the CACE estimation in a longitudinal setting. On one hand, longitudinal designs collect multiple data points on the same individual, thus more information can be used for class separation with the MMB estimation method. On the other, the LGM can also improve the estimation precision by isolating measurement errors for the outcome variable, which could also contribute to class separation. With the benefits of the LGM, it is questionable that the conclusion of the cross-sectional study would hold.

In addition, when noncompliance happens, there can be four types of subpopulations. Participants who always comply with their assignment are "compliers"; participants who always enroll themselves into the treatment condition regardless of their assignment are "always-takers"; individuals who always escape the treatment are "never-takers"; the last type is "defiers", meaning choosing the opposite of their randomly assigned levels. One assumption of CACE is that there are no defiers (Angrist et al., 1996; Angrist & Imbens, 1995); hence defiers were not discussed in this study

Past studies (Jo, 2002; Jo & Muthén, 2003), however, mainly included only two compliance statuses (complier and never-taker). Although some research designs are able to exclude always-takers by providing no access of treatment level to individuals assigned to the control group (e.g., Ialongo et al., 1999; Vinokur et al., 1995), it is not possible to always eliminate always-takers for all studies. For example, in the evaluation of the private school voucher program in Dayton, Ohio, Washington, D.C., and New York City by Howell and Peterson (2003), although participants were randomly selected to receive a voucher that was financially helpful for children from low-income families to attend private schools, many of those offered a voucher did not use it (i.e., never-takers) while a small portion who were not offered still chose private schools (i.e., always-takers). The same situation happened in the study of Angrist et al. (2002). In these scenarios, it is impossible and unethical to force students to attend a certain type of school, and for researchers, there are three compliance statuses. Therefore, it is necessary to examine the situation with three compliance statuses. By adding one more compliance group, the model becomes

more complicated, and previous research results using only two compliance groups may not be generalizable to this situation.

Last but not least, the well-known CACE estimation approach, the Standard IV approach, due to its simplicity in modeling and low computational burden, is also widely used in various studies. This approach can also adapt to the LGM framework to estimate longitudinal CACEs. Imbens and Rubin (1997b) demonstrated with a small simulation study with cross-sectional data that the Standard IV approach yielded more bias and less precision in estimating CACEs than the MMB approach. However, their simulation design only examined very limited conditions, and it is also unclear about how and to what extent the Standard IV approach would perform better or worse than the MMB approach for longitudinal CACE estimations.

Therefore, the purpose of the present study is to expand the literature of the longitudinal CACE estimation while using the LGM framework. This study is motivated by the fact that as the need to address the noncompliance issue in longitudinal experiments increases, there is not enough guidance for researchers and practitioners to decide on what the more important factors are regarding the research design and which estimation method (Standard IV vs. MMB based) they should choose. Specifically, this study aimed to answer four research questions:

1. Choosing from factors that previous studies have shown to influence the estimation of CACE, how will each of them affect the estimation success rate of the Standard IV and the MMB methods?

2. Among the six factors in Research Question 1, how will each affect the estimation accuracy (biased or unbiased estimation) of the Standard IV and the MMB methods?
3. When the effect size is not zero, among the six factors in Research Question 1, how will each affect the statistical power of the Standard IV and the MMB methods?
4. When the effect size is zero, among the other five factors in Research Question 1, how will each affect the empirical type I error rate of the Standard IV and the MMB methods?
5. Considering all criteria above, which method is recommended?

These questions were answered with findings from an intensive simulation design. The simulation was done by first generating datasets that differed in terms of the six investigated factors. The two estimation methods were then applied to each dataset respectively. In the end, the estimation results were aggregated for analysis and compared between the two estimation methods. With the analysis result, suggestions regarding experiment designs were provided for researchers and practitioners.

Chapter 2: Literature Review

2.1.Causal Inference

Almost in every *Introduction to Statistics* class, instructors always particularly emphasize the notion that *correlation does NOT imply causation*. A classic example to support this statement would be that in summer ice cream sales and crime rates are observed to be positively correlated (Salkind, 2010). It is obviously absurd to conclude that one causes the other, even without any statistical knowledge. A more rational explanation to this phenomenon could be that the pleasant temperature in summer increases the chance that people stay out late at night and it, therefore, increases the probability that burglaries, DUIs or other crimes occur. At the same time, high temperature also boosts the consumption of ice cream. Therefore, it is the temperature that causes the increase of ice cream sales and crime rates.

While acknowledging that correlation does not imply causation, one needs to understand the definition of cause, effect, and causal relationship. Shadish et al. (2002) used Mackie's definition of an inus condition –"an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition" (Mackie, 1974, p. 62) to describe most causes. In other words, there are many combinations of factors that can lead to the same effect. Each combination can independently cause the same effect. A single factor in that combination cannot lead to the effect unless there is only one factor in the combination, but each factor is necessary for that particular combination. Therefore, Shadish et al. (2002) described causal relationships as not deterministic but only increased the probability that an effect would occur.

The counterfactual model is widely used to describe an effect. Shadish et al. (2002) explained that it is only possible to observe what happens when people receive a condition level during an experiment. They describe that the counterfactual was the unobservable information of what "*would have happened*" (p. 5) to the same group of people if they had not received the condition in the experiment at the same moment. Therefore, an effect was the difference between what did happen and what would have happened. Holland (1986) summarized Rubin's model for causal inference (i.e., the RCM) with mathematical notations: the effect of a variable *T* on unit *i* measured by outcome *Y*, denoted as $Y_1(i)$, and relative to the effect of another variable *C*, denoted as $Y_0(i)$, was the difference, denoted as δ_i, between $Y_1(i)$ and $Y_0(i)$. If *T* is the treatment condition and *C* is the control condition, $Y_1(i)$ is the result of receiving the treatment level and $Y_0(i)$ is the result of receiving the control level for subject *i*. Only one result can be observed at the same moment; hence, one of $Y_1(i)$ and $Y_0(i)$ will become "what did happen" and the other will become the "what would have happened". The effect of the treatment condition relative to the control condition on subject *i* is the difference between $Y_1(i)$ and $Y_0(i)$,

$$\delta_i = Y_1(i) - Y_0(i). \quad (1)$$

2.2.Experiment

To establish a causal relationship between possible causes and effects, Shadish et al. (2002) summarized a theory by John Stuart Mill, a 19th-century philosopher, which states that a causal relationship existed if 1) the cause preceded

the effect, 2) the cause was related to the effect, and 3) no plausible alternative explanation for the effect can be found other than the cause.

Fortunately, an experiment would satisfy Mill's three conditions if it 1) manipulates the presumed cause and observes a forthcoming outcome, 2) checks whether the variation in the cause is correlated with the variation in the effect, and 3) uses feasible methods (e.g., random assignment, matching) during the experiment to reduce the plausibility of other explanations for the effect as well as ancillary methods to explore the plausibility of those could not be ruled out. Therefore, following Shadish et al. (2002), Murnane and Willett (2010) defined an experiment as "an empirical investigation" in which an independent "outside agent" had total manipulation on assigning cause levels to research participants and after the investigation, the consequences for an outcome would be measured (p. 30).

There are two types of experiments depending on how much control the "outside agent" has on the assignment of causal levels: in randomized experiments, the assignment is completely manipulated by researchers in order to achieve an entirely random assignment, and in quasi-experiments, researchers do not have full control over the assignment and as a result no randomness can be achieved. A well-executed randomized experiment would satisfy Mill's three conditions; therefore, a causal relationship could be easily inferred in randomized experiments.

2.2.1. Randomized experiments

According to Holland (1986), there is a "Fundamental Problem of Causal Inference" (p. 947)—it is impossible to observe the values of $Y_1(i)$ and $Y_0(i)$ on the same unit *i* at the same moment; therefore, it was impossible to observe the effect of

T on *i*. He provided two solutions for this problem: the scientific solution and the statistical solution.

The scientific solution was to exploit various homogeneity or invariance assumptions. Two methods could be used to apply the scientific solution. The first method assumed "*temporal stability*" (p. 948) and "*causal transience*" (p. 948). Temporal stability emphasized the constancy of response over time by assuming the value of $Y_0(i)$ did not depend on when the sequence of applying *C* to *i* first then measuring *Y* on *i* occurred. Causal transience asserted the effect of the cause *C* and the measurement process that resulted in $Y_0(i)$ was transient so it would not change *i* enough to influence $Y_1(i)$ measured later. A lot of physical devices would meet the two assumptions. A made up example could easily explain the notions. For example, if the effect of air pollution on the calculation speed of a computer is of interest, one can simply measure the computer calculation speed in an environment with clean air and then take the same computer to a room with polluted air. It does not matter when to conduct the first part of the experiment, so the temporal stability assumption is met. Plus, neither the clear air nor the measurement of computer calculation speed is likely to change the computer; therefore, the second measurement conducted in a polluted environment is not affected by the first measurement. In this way, the causal transience assumption is met.

The second method employed in the scientific solution only required unit homogeneity, that $Y_1(i_1) = Y_1(i_2)$ and $Y_0(i_1) = Y_0(i_2)$ for two units i_1 and i_2. Under this assumption the causal effect of *T* was $Y_1(i_1) - Y_0(i_2)$. Take the previous

computer experiment as an example. If the researcher is convinced that the same type of computers would not differ in all relevant aspects, calculation speed or reaction to polluted air, for instance, the unit homogeneity is met and the causal effect of polluted air is the difference in calculation speeds for *two* computers measured separately in two conditions.

However, in most experiments, neither method of the two scientific solutions would be applicable. That is when the statistical solution is engaged. Holland (1986) stated that the statistical solution tried to estimate the average causal effect, Δ, of *T* (relative to *C*) over a population *I*, which is equal to the expected value of the difference, $Y_1(i) - Y_0(i)$, over every unit *i* in *I*. This is defined as:

$$\Delta = E(Y_1 - Y_0) = E(Y_1) - E(Y_0). \tag{2}$$

It is worth noticing that for a given unit *i* only one of $Y_1(i)$ or $Y_0(i)$ could be observed based on *i*'s value on variable *D*. If *i* was assigned and attended to the treatment group *T*, then $D = 1$; otherwise, $D = 0$. Therefore, data $\left(D(i), Y_D(i)\right)$ for unit *i* could be observed. It was also known that

$$E(Y_D \mid D = 1) = E(Y_1 \mid D = 1), \tag{3}$$

and

$$E(Y_D \mid D = 0) = E(Y_0 \mid D = 0). \tag{4}$$

In general, $E(Y_1 \mid D = 1) \neq E(Y_1)$ because $E(Y_1 \mid D = 1)$ was the average value of $Y_1(i)$ over only those exposed to treatment but $E(Y_1)$ the average value of $Y_1(i)$ overall *i* in *I*. In other words, the former was the averaged value of a subpopulation while the later was the averaged value of a whole population. If the subpopulation

differs from the whole population on factors that may influence the outcome, their expectations of the outcome should also differ. Similarly, $E\left(Y_0 \mid D=0\right) \neq E\left(Y_0\right)$. However, when randomized experiments were used, the two subpopulations are randomly formed by randomly dividing the whole population, so the two subpopulations are not expected to deviate from the whole pupation. Accordingly, their expectation of the outcome under the treatment or control condition should be the same as the full population's expectations under the two conditions. In equation,

$$E(Y_1 \mid D=1) = E\left(Y_1\right), \tag{5}$$

and

$$E(Y_0 \mid D=0) = E\left(Y_0\right) \tag{6}$$

when random assignment is achieved. Therefore, the average casual effect Δ could be expressed as

$$\Delta = E(Y_1 \mid D=1) - E(Y_0 \mid D=0) \tag{7}$$

2.2.2. Randomized experiment with noncompliance.

Unfortunately, randomized experiments cannot always be perfectly implemented. In some cases, although researchers can independently manipulate a potential cause and try to draw counterfactual inference about what would have happened in the absence of treatment, the random assignment process could be easily contaminated. Shadish et al. (2002) exemplified that non-randomness could happen because of self-selection, i.e., participants choosing which treatment group they wanted to join, or as a result of administrator selection, where, instead of random assignment, teachers, legislators, therapists, or others parties chose which participants

to receive certain treatments. Consequently, the randomization is invalidated by the noncompliance.

Random experiments with noncompliance meet the first two of Mill's three conditions for making causal inferences: 1) cause precedes effect, and 2) different levels of cause can lead to different levels of effects. However, the third condition of using random assignment to eliminate all other possible explanations for any differences in effect observed is violated by noncompliance. The control group may differ from the treatment group in many systematic ways other than the manipulated treatment conditions.

In randomized experiments, the assignment of participants to treatment and the treatment level received are exogenous because the "outside agent" has total control over the assignment and receipt to ensure the randomness. With noncompliance, researchers have no direct control over what level of treatment is received by subjects. Therefore, the treatment received is considered to be endogenous. As a result, Equations 5 and 6 will not hold; therefore the average causal effect cannot be directly estimated using Equation 7.

As discussed in the earlier section, the special version of the IV approach, the CACE approach, can be used with the occurrence of noncompliance. The IV method is regarded as the "most powerful weapon" (Angrist & Pischke, 2008, p. 114) among tools of estimating the causal effect. The exogenous instrumental variable can carve out the exogenous variation in the independent variable and estimate the causal effect on the outcome variable with only this part (Murnane & Willett, 2010). Therefore, an

"asymptotically unbiased estimate" of causal effect can be estimated even with an endogenous observed treatment variable (Murnane & Willett, 2010, p. 205).

2.2.3. Compliers, always-takers, never-takers, and defiers

In order to understand the mechanism of using the CACE approach to estimate the causal effect, the population of interest should be clearly defined first. Imbens and Rubin (1997b) characterized study participants as compliers, always-takers, never-takers, and defiers. Compliers will always follow the assignment. In other words, if a complier is assigned to the treatment group, he or she will take the treatment, and if this complier is assigned to the control group, he or she will not take the treatment. Always-takers will always make themselves available for the treatment and take the treatment no matter whether they are assigned to treatment or control group. Never-takers, on the other hand, will always avoid taking the treatment regardless of their group membership. The last type of participant, defiers, will choose to take the different treatment that is the opposite of their own assignment: a defier being assigned to the treatment group will not take the treatment but will take the treatment once he or she is assigned to the control group.

Let Z_i be the binary assignment variable, for participant *i*, taking the value of 1 if the *i*th participant is randomly assigned to the treatment group and 0 if randomly assigned to the control group. Let $D_i = D_i(Z_i)$ that will take two forms, $D_i(1)$ and $D_i(0)$. $D_i(1)$ and $D_i(0)$ denote the treatment level received or observed giving the random assignment of *Z*. If participant *i* is a complier, $D_i(1) = 1$ and $D_i(0) = 0$. If participant *i* is an always-taker, $D_i(1) = D_i(0) = 1$. If participant *i* is a never-

taker, $D_i(1) = D_i(0) = 0$. If participant *i* is a defier, $D_i(1) = 0$ and $D_i(0) = 1$.

In most cases, the values for Z_i and $D_i(Z_i)$ are observable, but the *i*th participant's compliance status is usually unknown. For example, if a participant is assigned to the treatment group, $Z_i = 1$, and he actually takes the treatment, $D_i(1) = 1$, he can be a complier or an always-taker. Or if a participant is assigned to the control group, $Z_i = 0$, but he actually takes the treatment, $D_i(0) = 1$, he can be a never-taker or a defier. Because of this unknown compliance status, the estimation of the treatment effect becomes more complicated.

2.2.4. Estimating causal effect with complete compliance

If all participants comply with their assignments (i.e., there is no noncompliance), our experiment would be the typical randomized experiment where the causal effect of the treatment equals the intent-to-treat (ITT) effect, and

$$\Delta = ITT = \mu_1 - \mu_0 \ , \tag{8}$$

where μ_1 denoted the population mean of those who were assigned to the treatment group and μ_0 denoted the population mean of those who were assigned to the control group. Under the assumption of stable unit treatment value (SUTVA), which requires no correlation between the value of potential outcome for each participant and the treatment status of other participants in the sample (Rubin, 1974, 1980, 1990) and the assumption of random assignment of participants into different treatment groups, the unbiased estimate of the causal effect with full compliance is

$$\Delta = ITT = \overline{Y_1} - \overline{Y_0}, \quad (9)$$

where $\overline{Y_1}$ is the sample mean outcome of participants assigned to the treatment group, and $\overline{Y_0}$ is the sample mean outcome of participants assigned to the control group. The SUTVA assumption and the random assignment assumption is explained in section 2.2.5.

To describe the ITT effect, consider a hypothetical research scenario where researchers are interested in the effect of a new teaching method on improving students' performance on a reading test. Students are randomly assigned to take the new and the old teaching method without knowing which group takes the new method and which group takes the old method. Everything else is also kept the same for both groups. After a certain time, all participants are tested with the same reading test. If every student in the study follows the assignment, the difference between the group mean scores is an unbiased estimation of the treatment effect of the new teaching method. However, if there is noncompliance, the group mean difference is just an unbiased estimation of the assignment effect, i.e., the ITT effect.

2.2.5. Estimating causal effect with noncompliance

As discussed above, there can be compliers, always-takers, never-takers, and defiers in a sample, and each group of participants could have its own distribution of causal effect. When there is no noncompliance and it is feasible to create control and treatment groups that are expected to have the same proportion of each of the four subpopulations, the treatment effect can easily be estimated with the difference between the treatment and the control groups. When noncompliance is involved, the

difference between the group of subjects who are assigned to take the treatment and those who are assigned to take the control is merely the intent-to-treat effect that reflects the influence of assignment instead of treatment because there are no counterfactual groups in the sample for some compliance groups.

If, however, it is possible that the researcher can force all subjects assigned to the treatment group to take the treatment (e.g., injection of a new vaccine) and leaves no access to the treatment for those assigned to the control group, the ITT effect is still the causal effect of this treatment on the whole population despite the underlying compliance statuses. The enforcement eliminates self-selection for the non-complier subpopulations. The effect estimated by the group difference is an unbiased estimation of the treatment effect overall subpopulations.

However, the mixed treatment effect overall subpopulations might not be the focal effect of interest. For example, in an educational intervention study, policymakers are interested to know the effect of a type of tutorial software on students' achievement under natural settings. Once out of the enforced experiment, students will be free to choose whether they want to use the software or not and there may be different averaged effects for each subpopulation. Therefore, the estimated causal effect observed by forcing every subject in the sample to comply with his or her own assignment cannot be generalized to the effect of interest. This is a downside associated with a forced compliance design.

In addition, the situation of forced treatment and restricted access to treatment for the control group is very rare. In a lot of research settings, researchers can at most randomly assign subjects to the treatment or control group, but they have no further

control over whether subjects will actually comply with the assignment. After participants are randomly assigned to the treatment or control group, some subjects will comply with their assignment and some subjects will not comply. Table 1 illustrates the composition of each group with the presence of noncompliance.

Table 1

Possible Composition of the Compliance Status for a Sample with Noncompliance

		Actual Treatment Taking (D)	
		1 (Take Treatment)	0 (Not Take Treatment)
Assignment (Z)	1 (Treat)	Cell 1: Compliers Always-takers	Cell 2: Never-takers Defiers
	0 (Control)	Cell 3: Always-takers Defiers	Cell 4: Compliers Never-takers

The RCM (Holland, 1986; Rubin, 1974, 1980) explains the difference between the forced design and the design where the enforcement is not possible. According to RCM, every individual has two possible outcomes $Y_1(i)$, if taking the treatment, and $Y_0(i)$, if taking the control assignment, but only one outcome can be observed and the other outcome is intrinsically missing. Table 2 presents the observed and missing situations for both designs. Outcomes that are not formatted are observed outcomes, and outcomes that are underlined and shaded in grey are missing.

For the forced compliance design, subjects assigned to the treatment group all miss their possible outcomes for the control level and subjects assigned to the control group all miss their possible outcomes for the treatment level. As subjects were randomly assigned to the treatment or control group, the two groups are expected to be equivalent on the outcome distributions. Therefore, the control group mean of

$Y_0(i)$ is an unbiased estimation of the mean of the missing values on $Y_0(i)$ in the treatment group.

On the contrary, for the non-forced compliance design, only compliers and defiers have both observations, $Y_1(i)$ and $Y_0(i)$, across the two assigned groups. For always-takers and never-takers, neither potential outcome of the two levels of the assignment is observable. Therefore, it is impossible to estimate an unbiased treatment effect for always-takers and never-takers. As illustrated in Table 2, for always-takers, no matter which group they are assigned to, only $Y_1(i)$ is observable. Similarly, for never-takers, only $Y_0(i)$ is observable. Thus, the ITT effect is not an averaged causal effect over all subpopulations in non-forced compliance designs.

Table 2

Outcome Compositions for the Forced Compliance Design and the Non-Forced Compliance Design

		Forced Compliance Actual Treatment Taking						Non-Forced Compliance Actual Treatment Taking					
		1			0			1			0		
Assignment	1	Compliers	$Y_1(i)$	$\underline{Y_0(i)}$				Compliers	$Y_1(i)$	$\underline{Y_0(i)}$			
		Always	$Y_1(i)$	$\underline{Y_0(i)}$				Always	$Y_1(i)$	$\underline{Y_0(i)}$			
		Never	$Y_1(i)$	$\underline{Y_0(i)}$							Never	$\underline{Y_1(i)}$	$Y_0(i)$
		Defiers	$Y_1(i)$	$\underline{Y_0(i)}$							Defiers	$\underline{Y_1(i)}$	$Y_0(i)$
	0				Compliers	$\underline{Y_1(i)}$	$Y_0(i)$				Compliers	$\underline{Y_1(i)}$	$Y_0(i)$
					Always	$\underline{Y_1(i)}$	$Y_0(i)$	Always	$Y_1(i)$	$\underline{Y_0(i)}$			
					Never	$\underline{Y_1(i)}$	$Y_0(i)$				Never	$\underline{Y_1(i)}$	$Y_0(i)$
					Defiers	$\underline{Y_1(i)}$	$Y_0(i)$	Defiers	$Y_1(i)$	$\underline{Y_0(i)}$			

Note. Outcomes underlined and shaded in grey are missing

One might be curious if it is conceivable to use the group difference between the group that actually takes the treatment (i.e. the $D = 1$ group) and the group that takes the control level (i.e. the $D = 0$ group) to estimate the treatment effect.

Unfortunately, the two groups are not equivalent. When noncompliance occurs, the $D = 1$ group has no never-takers while the $D = 0$ group has no always-takers. The intended group balance for using random assignment is not achieved in this situation. Therefore, using the group difference as the estimated treatment effect is not defendable.

Fortunately, the compliers still have a balanced distribution of $Y_1(i)$ and $Y_0(i)$ across the two assigned groups. If it is somehow feasible separate the complier subpopulation, the treatment effect on the compliers (the CACE) can still be estimated. The IV method, with some restrictions, makes it possible. The CACE can be defined as

$$\Delta_c = CACE = \mu_{1c} - \mu_{0c} \tag{10}$$

where μ_{1c} denotes the population mean of compliers that have been assigned to the treatment group and μ_{0c} denotes the population mean of compliers that have been assigned to the control group.

Assumptions when using the Standard IV method to estimate the CACE. It might seem easy to estimate CACE with Equation 10, but it is almost impossible to verify some participants' compliance status. For example, in Cell 1 of Table 1 where Assignment (*Z*) takes 1 and Actual Treatment Taking (*D*) also takes 1, it is clear that participants who are assigned to take the treatment and as a result take the treatment, but it is implausible to distinguish, without further information, whether they just take the treatment because they are assigned to the treatment group (a complier) or whether they will take the treatment despite their membership (an always-taker). Likewise, in Cell 2 where Assignment (*Z*) takes 1 and Actual Treatment Taking (*D*)

takes 0, one can know that participants in the cell are assigned to take the treatment but they fail to show up, and one can hardly sort out whether they just refuse to take any treatment (a never-taker) or they only take the opposite of their assignment (a defier). In fact, for a sample that has the exact composition shown in Table 1, it is impossible to identify any participant's compliance status without additional information.

Nonetheless, following the work of Angrist et al. (1996) and Angrist and Imbens (1995), five assumptions based on the RCM put more restrictions on the data and allows us to estimate the CACE.

The first assumption, SUTVA, requires that potential outcomes for each individual should not be related to the treatment status of other subjects in the sample. An example in educational intervention could be an experiment that looks into the effect of class size on students' academic performance. Students are assigned to two experiment groups, the treatment group, a smaller sized class, and the control group, a normal sized class. However, there are two good friends in the sample. Let's call them Jack and James. Jack will perform particularly well if James is in the same classroom, and he will not focus enough if otherwise. As a result, the performance of Jack is not only influenced by his assignment but also by the presence of James. If there are a lot of individuals similar to Jack in the sample, the treatment effect is further intertwined with the codependence among sample units and becomes impossible to estimate.

The second assumption requires random assignment of subjects to the treatment and the control group. That is to say, the probability of being assigned to

$Z_i = 1$ should be equal across all subjects and the same goes for the probability of being assigned to $Z_i = 0$. Random assignment ensures that the assignment variable, *Z*, is exogenous and approximates the equivalence between the treatment and the control conditions. If random assignment is successfully done, the pretreatment characteristics between the two groups will be very similar, which provides a fair comparison between the two groups.

As mentioned before, the assumption of SUTVA and random assignment ensures that the difference between the control and the treatment groups in the sample is an unbiased estimation of the ITT effect in the population. There are three more assumptions needed for an unbiased estimation of the CACE.

The third assumption is the assumption of monotonicity: there are no defiers in the population. This assumption will hold if participants have no option other than to take their assigned treatment or there is no reason for someone to be a defier. The first condition is equivalent to the forced treatment mentioned above, whereas the second condition needs convincing and comprehensive arguments that rule out all possibilities of having defiers in the population. According to Angrist and Imbens (1995), the assumption of monotonicity is "fundamentally untestable" (p. 469) and the validity needs to be argued in a specific context. In most cases, defiers are considered as the least likely type of noncompliance, but the violation of this assumption can have a serious impact on the estimation of CACE using the IV method (Jo, 2002).

Assumption 4 is the assumption of exclusion restriction: for never-takers and always-takers, the distributions of the potential outcome variable *Y* are not related to

the treatment assignment variable *Z*, which captures the notion that any effect of *Z* on *Y* must be via an effect of *Z* on *D*. In other words, never-takers and always-takers will always receive their preferred treatment irrespective of the original assignment (Z_i): D_i always equals 1 for always-takers and always equals 0 for never-takers.

The last assumption requires a nonzero-average causal effect of *Z* on *D* or the expected difference between the proportion of participants who actually receive treatment assignment and taking the treatment ($Z_i = 1$ and $D_i = 1$) and the proportion of participants who actually receive control group assignment but taking the treatment ($Z_i = 0$ but $D_i = 1$) is not zero (Little & Yau, 1998). More explanation is provided in the next section.

2.3.Instrumental Variable

2.3.1. Standard IV estimation

With the five assumptions listed above, one can use the Standard IV estimation method to estimate the averaged causal effect among compliers. Following Imbens and Rubin (1997b) and Little and Yau (1998), this section demonstrates the mathematical derivation for the CACE estimation process with the Standard IV method.

In Equation 8, the ITT effect is the difference between the population means of those assigned to the treatment group and those assigned to the control group. It can also be decomposed as a combined causal effect of the assignment *Z* on compliers and non-compliers,

$$\begin{aligned}\Delta = ITT &= \mu_1 - \mu_0 \\ &= \pi_c \Delta_c + \pi_{nc} \Delta_{nc} \\ &= \pi_c \Delta_c + (1 - \pi_c) \Delta_{nc},\end{aligned} \tag{11}$$

where π_c is the proportion of compliers in the population, π_{nc} is the proportion of non-compliers, Δ_c is the averaged causal effect of the assignment on compliers (i.e. the CACE) and Δ_{nc} is the averaged causal effect of the assignment on all non-compliers—defiers, always-takers, and never-takers. As the population consists of compliers and non-compliers, the sum of π_c and π_{nc} should equal 1. Equation 11 can be further changed:

$$\Delta_c = \frac{\Delta - (1 - \pi_c)\Delta_{nc}}{\pi_c}. \tag{12}$$

Δ_{nc} can be decomposed into a combination of the averaged causal effects of the assignment on never-takers, always-takers, and defiers respectively,

$$\Delta_{nc} = \pi_{at} \Delta_{at} + \pi_{nt} \Delta_{nt} + \pi_d \Delta_d, \tag{13}$$

where π_{at}, π_{nt}, and π_d are the population proportions of always-takers, never-takers, and defiers respectively, Δ_{at}, Δ_{nt}, and Δ_d are the averaged causal effects of the assignment on always-takers, never-takers and defiers correspondently.

According to assumption 3, monotonicity, there are no defiers in the population. π_d hence equals to 0. With the exclusion restriction, which assumes the potential outcome variable *Y* has distributions not related to the treatment assignment variable *Z* for never-takers and always-takers, the causal effects of the assignment on never-takers and always-takers are 0. Therefore, $\Delta_{at} = \Delta_{nt} = 0$ as described below.

Figure 1 illustrates more of the exclusion restriction assumption by specifying the path model for each subpopulation when this assumption is met. The four numbers from *Z* to *D* denote the effects of the assignment variable on the treatment-received variable for the four subpopulations. The different βs from *D* to *Y* are the effects of the treatment-received variable on the outcome for the four subpopulations. Correspondingly, the products of the two numbers are the effects of the assignment variable on the outcome (i.e., Δ_c, Δ_{at}, Δ_{nt}, and Δ_d). As both *Z* and *D* are dichotomous variables, 1 from *Z* to *D* means *Z* and *D* always take the same level (i.e., compliers), -1 from *Z* to *D* means *Z* and *D* always take the opposite level (i.e., defiers), and 0 from *Z* to *D* means the value of *D* is not related to the value of *Z* (i.e., always-takers and never-takers). As a result, the causal effect of *Z* on *Y* for compliers is $\Delta_c = 1 \times \beta_c$. For defiers is $\Delta_d = -1 \times \beta_d$. Because monotonicity assumes no defiers, this part will not be added to Δ. For always-takers, the causal effect of *Z* on *Y* is $\Delta_{at} = 0 \times \beta_{at} = 0$. Similarly, for never-takers, the causal effect of *Z* on *Y* is $\Delta_{nt} = 0 \times \beta_{nt} = 0$. Subsequently, with assumption 3 and 4, Equation 13 equals to 0 and Equation 12 can be simplified as

$$\Delta_c = \frac{\Delta}{\pi_c}. \tag{14}$$

As mentioned above, $\overline{Y}_1 - \overline{Y}_0$ is an unbiased estimation of Δ under the assumption of SUTVA and random assignment. Thus, the next step is to find an unbiased estimation of π_c.

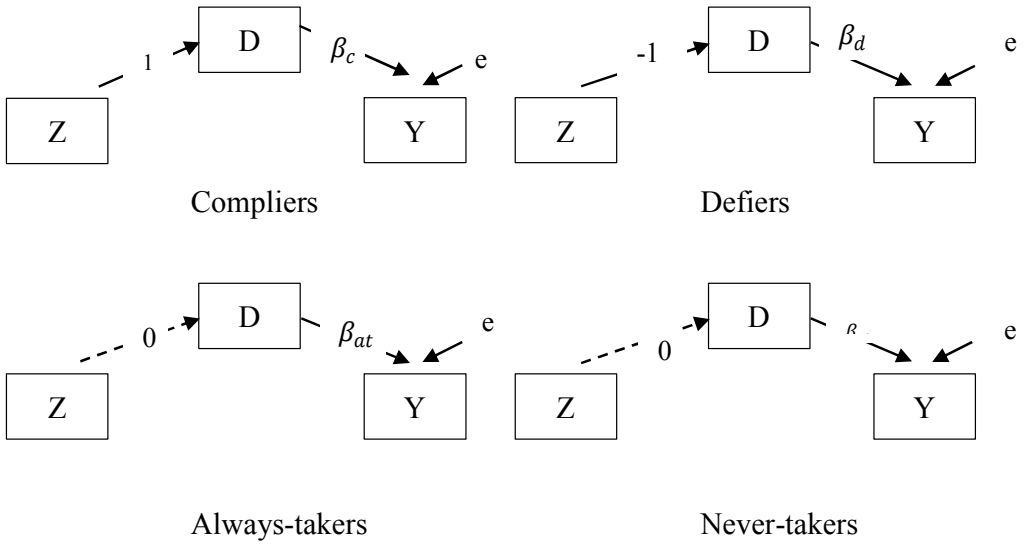

Figure 1. Model specifications for the four subpopulations.

The proportion of compliers is equal to the proportion of compliers and always-takers minus the proportion of always-takers,

$$\pi_c = \pi_{c+at} - \pi_{at}, \tag{15}$$

where π_{c+at} is the proportion of compliers and always-takers. Let p_{c+at} be the proportion of participants who are assigned to the treatment group and actually take the treatment, i.e., the proportion of compliers and always-takers in the treatment group, and p_{d+at} be the proportion of participants who are assigned to the control group but instead take the treatment, i.e., the proportion of defiers and always-takers in the control group. Due to the assumption of monotonicity, $p_d = 0$, so $p_{d+at} = p_{at}$. Under assumption 2, p_{c+at} is an unbiased estimation of π_{c+at} and p_{at} is an unbiased estimation of π_{at}. Thereafter, an unbiased estimation of Δ_c, denoted as d_c, can be expressed as

$$d_c = \Delta_c = \frac{\Delta}{\pi_c} = \frac{\Delta}{\pi_{c+at} - \pi_{at}} = \frac{\bar{Y}_1 - \bar{Y}_0}{p_{c+at} - p_{at}} \tag{16}$$

As all participants are randomly assigned to take either $Z_i = 1$ or $Z_i = 0$, π_{c+at} and π_{at} are expected to be equal in both $Z_i = 1$ or $Z_i = 0$ groups. Therefore, the π_{c+at} for the $Z_i = 1$ group and the π_{at} for the $Z_i = 0$ group are also equal to the π_{c+at} and π_{at} in the whole population, so the p_{c+at} for the $Z_i = 1$ group and the p_{at} for the $Z_i = 0$ group are an unbiased estimation of population π_{c+at} and π_{at}. The equation above can be further decomposed as

$$d_c = \frac{\bar{Y}_1 - \bar{Y}_0}{p_{c+at} - p_{at}} = \frac{\frac{\sum_{i\in(Z=1)} Y_i}{n_{(Z=1)}} - \frac{\sum_{i\in(Z=0)} Y_i}{n_{(Z=0)}}}{\frac{n_{(Z=1,D=1)}}{n_{(Z=1)}} - \frac{n_{(Z=0,D=1)}}{n_{(Z=0)}}}, \tag{17}$$

where, as presented in Table 1, $n_{(Z=1)}$ is equal to the sample size of Cell 1 plus Cell 2, $n_{(Z=0)}$ is equal to the sample size of Cell 3 plus Cell 4, $n_{(Z=1,D=1)}$ is equal to the sample size of Cell 1, and $n_{(Z=0,D=1)}$ is equal to the sample size of Cell 3. The unbiased estimation of the causal effect of the treatment on the compliers is equal to the ITT effect divided by the difference in the proportion of compliers and always-takers in the assigned treatment group and the proportion of always-takers in the assigned control group.

Assumption 5 requires a nonzero-average causal effect of *Z* on *D*, which is equivalent of expecting the difference between the proportion of participants who

actually receive treatment assignment and taking the treatment and the proportion of participants who actually receive control group assignment but taking the treatment to be not zero. As discussed above, the causal effect of *Z* on *D* can be decomposed into products of the effect of each subpopulation times each subpopulation's proportion. In addition, the effects for the always-takers and never-takers are zero and the effect for compliers is 1. It is further assumed there are no defiers. Therefore, in order to have a nonzero-average causal effect of *Z* on *D*, the proportion of the compliers should not be zero. On the other hand, as demonstrated in the above paragraph, the proportion of the compliers can be estimated by the difference between the proportion of participants who actually receive treatment assignment and taking the treatment and the proportion of participants who actually receive control group assignment but taking the treatment; thus when the proportion difference is not zero, the proportion of the compliers is not zero and the average causal effect of *Z* on *D* is not zero.

Although Assumption 5 requires that $\pi_c = \pi_{c+at} - \pi_{at}$ to be not zero and π_c should be by definition nonnegative, the unbiased estimation

$p_c = p_{c+at} - p_{at} = \frac{n_{(Z=1,D=1)}}{n_{(Z=1)}} - \frac{n_{(Z=0,D=1)}}{n_{(Z=0)}}$ cannot be guaranteed to be bigger than zero.

This is one drawback of using the Standard IV approach.

There are different variants of the Standard IV estimation. When both the instrumental variable and the assignment variable are dichotomous, the Standard IV estimate presented in Equation 17 is called the Wald estimator (Brookhart, Rassen, & Schneeweiss, 2010; Murnane & Willett, 2010). A more general estimation method is the method-of-moments estimator, where all variables can be either dichotomous or

continuous (Murnane & Willett, 2010). The method-of-moments estimator can be expressed as

$$d_c = \frac{s_{YZ}}{s_{DZ}}, \tag{18}$$

where s_{YZ} is the covariance between *Y* and *Z*, and s_{DZ} is the covariance between *D* and *Z*. Equation 18 is equivalent to Equation 17 when variable *D* and *Z* are dichotomous.

When a study involves only one outcome variable, one assignment-taken variable, and one random assignment variable, both the Wald estimator and the method-of-moments estimator work well (Murnane & Willett, 2010). However, with more complex research designs, the two-stage least-square (2SLS) or the structural equation modeling (SEM) method are widely used. The two methods can easily accommodate various data-analytic settings, such as including multiple random assignments, assignment-taken, and outcome variables, and incorporating non-linear relationships among variables (Brookhart et al., 2010; Murnane & Willett, 2010).

Modeling with a continuous assignment-taken variable. Firstly, consider when the assignment-taken and outcome variables are both continuous. The 2SLS method is a split process that uses two ordinary least square (OLS) regressions to estimate the treatment effect among compliers. The first stage uses the exogenous assignment variable, *Z* (it will be a vector symbol, ***Z***, if there are more than one assignment variables), to carve out the exogenous part in the treatment-taken variable, *D*:

$$D = \alpha_0 + \alpha_1 Z + e_1 \tag{19}$$

The second stage then uses only the carved out exogenous part from stage 1 ($\hat{D} = \hat{\alpha}_0 + \hat{\alpha}_1 Z$) to predict the outcome variable, *Y*:

$$Y = \beta_0 + \beta_1 \hat{D} + e_2 = \beta_0 + \beta_1 \times (\hat{\alpha}_0 + \hat{\alpha}_1 Z) + e_2, \quad (20)$$

where β_1 is the estimated treatment effect among compliers.

The SEM method fits Equation 19 and Equation 20 simultaneously and yields identical parameter estimates as the 2SLS method. Nonetheless, the SEM approach has two main advantages over the 2SLS method. The first advantage is that the SEM method depicts the relationships among all variables in a clearer way. For example, suppose there is one assignment variable *Z*, one assignment-taken variable *D*, and one outcome variable *Y*. If all assumptions discussed earlier are met, the path model is presented in Figure 2. There is no arrow pointing to the exogenous variable *Z*, which is consistent with the aforementioned assumption of random assignment. Also, there is no connection between *Z* and *Y* other than through *D*, which conforms to the assumption of exclusion restriction. In order to meet the last assumption of nonzero-average causal effect of *Z* on *D*, α_1 should not be 0. Parameters indicating causal effects in Figure 2 (α_1 and β_1) are identical to parameters estimated from Equation 19 and 20, so β_1 is still the estimated treatment effect among compliers.

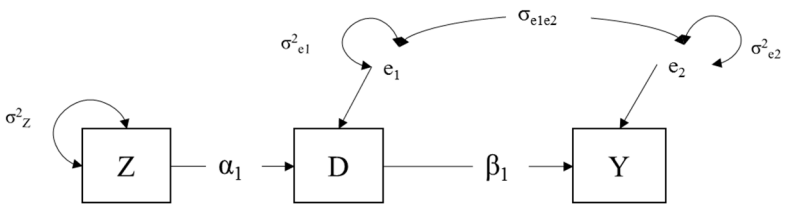

Figure 2. Path model presenting hypothesized relationships among an outcome variable *Y*, an endogenous assignment-taken variable *D*, and an exogenous assignment variable *Z*.

Note that the first stage error, e_1, is allowed to covary with the second stage error, e_2. This specification is essential because it separates the variation that is not "caused" by *Z* in *D* and *Y* from the variation that is "caused" by *Z*. In this way, the exogenous part in *D* can be utilized to determine its effect on *Y* (i.e., β_1) (Murnane & Willett, 2010). By conducting a hypothesis test of the existence of the covariation between e_1 and e_2, it can be decided if the instrumental variable *Z* is required to estimate the causal effect of *D* on *Y*: if the covariation is not significant, the dependent variable *Y* can be simply regressed on *D*.

The second advantage of the SEM method is that the outcome variable can be measured with multiple indicators and only the relatively "pure reflection of the variable of interest" (Hoyle, 2012, p. 12) will be used for treatment effect estimation. In this way, the outcome variable can be used as a latent factor and measurement errors from indicators can be adequately eliminated. This feature can easily be applied to LGM techniques where each measurement occasion can be treated as an indicator of the latent intercept and latent slope. In the current research, the author only focused on the SEM variant.

Modeling with a dichotomous assignment-taken variable. If *D* is continuous, conventional SEM approach with common SEM packages (e.g., Amos [Arbuckle, 2006], EQS 6 [Bentler, 2006], LISREL 9.2 [Jöreskog & Sörbom, 2015], Mplus 8.1 [Muthén & Muthén, 1998-2018], the open source R package 'lavaan' [Rosseel, 2012]) are available to estimate all parameters in the path model and a *z*-test allows the examination of the significance of each parameter. However, when *D* is dichotomous, treating *D* as continuous will yield an estimation of β_1 that is identical to the Wald estimator. To illustrate the computation of β_1, a covariance matrix of the three observed variables for the path model is presented in Figure 3 according to Wright's rule of tracing (1918, 1934). β_1 can be calculated directly by dividing the covariance of *Z* and *Y* by the covariance of *Z* and *D*, which is equivalent to Equation 18. As stated earlier, Equation 17 and 18 are equivalent with dichotomous *Z* and *D*. Therefore, using conventional SEM and treating variable *D* as continuous will result in an asymptotically unbiased estimation of the CACE when all other assumptions are met.

$$\begin{array}{c} \\ Z \\ D \\ Y \end{array} \begin{array}{c} \begin{array}{ccc} Z & D & Y \end{array} \\ \begin{bmatrix} \sigma_Z^2 & & \\ \alpha_1\sigma_Z^2 & \alpha_1^2\sigma_Z^2+\sigma_{e_1}^2 & \\ \alpha_1\beta_1\sigma_Z^2 & \left(\alpha_1^2\sigma_Z^2+\sigma_{e_1}^2\right)\beta_1+\sigma_{e_1e_2} & \left(\alpha_1^2\sigma_Z^2+\sigma_{e_1}^2\right)\beta_1^2+2\beta_1\sigma_{e_1e_2}+\sigma_{e_2}^2 \end{bmatrix} \end{array}$$

Figure 3. Covariance matrix of variable *Z*, *D,* and *Y*.

However, the standard errors estimated would be inaccurate because conventional SEM requires continuous dependent variables (Edwards, Wirth, Houts, & Xi, 2012) that are multivariate normal and have homoscedastic residuals (Kline,

2012). With the binary assignment-taken variable, none of the assumptions are met. Hence, the standard errors estimated by directly using the SEM model would be inaccurate and therefore lead to incorrect rejection or retain. Fortunately, bootstrapped standard errors are available for an empirical estimation of the standard error. Poi (2004) presented a clear illustration of this nonparametric technique. If the research interest was to estimate a parameter θ for population F and the current sample f of size n was randomly selected from F, the sample data can be used to obtain $\hat{\theta}$ as an estimation of θ. With bootstrapping, one can repeatedly draw random samples f_i of the size n with replacement from the original sample f, and from each randomly drawn parameter $\hat{\theta}_i$ can be estimated. After repeating this B times, all $\hat{\theta}_i$s will form an empirical sampling distribution for $\hat{\theta}$. With the sampling distribution, the standard deviation of the distribution can serve as the standard error of $\hat{\theta}$. With higher repetition, the bootstrap sampling distribution will approximate the real sampling distribution better (Poi, 2004); however, the total computation time will increase a lot. In this study, the repetition time was set to be 500 (B = 500) as a compromise between computation time and standard error precision.

2.3.2. Mixture model based method for the CACE estimation

The Standard IV estimation method can give a causal interpretation for the compliers without requiring a functional form or constant treatment effect assumptions (Imbens & Rubin, 1997b). However, its merit of being simple and clean also leads to some limitations.

Imbens and Rubin (1997b) laid out the limitation of the Standard IV method and pointed out that the estimates were essentially based on the estimation of the density function of the outcome variable for the compliers. However, during the estimation process, the density function was not restricted to be positive, so the estimation results can be inaccurate. Their detailed explanations are provided below.

Let $f_{zd}(Y)$ denote the distribution of the outcome Y_i in a subpopulation where the individuals take $Z_i = z$ and $D_i = d$. $f_{zd}(Y)$ can be estimated with the observable empirical distribution $\hat{f}_{zd}(Y)$ from a subsample where the subjects take $Z_i = z$ and $D_i = d$. Let $g_{kz}(Y)$ denote the distribution of the outcome Y_i in a subpopulation where individuals take $K_i = k$ and $Z_i = z$. K is a latent variable describing subjects' real compliance status and K_i can take on four possible values as follows,

$$K_i = \begin{cases} c, \text{ if subject } i \text{ is a complier} \\ d, \text{ if subject } i \text{ is a defier} \\ at, \text{ if subject } i \text{ is an always-taker} \\ nt, \text{ if subject } i \text{ is a never-taker} \end{cases}.$$

As monotonicity assumes that there are no defiers, K_i would not take the value *d*. Because K_i is not directly observable for all individuals (i.e. one cannot distinguish compliers from always-takers in the treatment group and cannot distinguish compliers from never-takers in the control group either), the empirical distributions, $\hat{g}_{kz}(Y)$, for all subgroups cannot be obtained and $g_{kz}(Y)$ is not directly estimable for every subgroup.

However, the estimable distribution of $f_{zd}(Y)$ is available for deriving the distribution of $g_{kz}(Y)$. Because of randomization, the assignment variable *Z* is independent of subjects' true compliance status *K*. Also because of exclusion restriction, the distributions for always-takers and never-takers on the potential outcome variable *Y* are the same irrespective of which value subjects take for *Z*. If the sample size is big enough, the observed empirical distribution of Y_i for always-takers assigned to the control group, $\hat{f}_{01}(y)$, enables the estimation of the distribution of Y_i for always-takers in both assignment groups. The same logic also applies to the never-takers. Therefore,

$$g_{at}(y) = g_{at1}(y) = g_{at0}(y) = f_{01}(y), \tag{21}$$

$$\hat{g}_{at}(y) = \hat{g}_{at1}(y) = \hat{g}_{at0}(y) = \hat{f}_{01}(y), \tag{22}$$

and

$$g_{nt}(y) = g_{nt1}(y) = g_{nt0}(y) = f_{10}(y), \tag{23}$$

$$\hat{g}_{nt}(y) = \hat{g}_{nt1}(y) = \hat{g}_{nt0}(y) = \hat{f}_{10}(y). \tag{24}$$

For compliers assigned to the control or the treatment group, their distributions on the outcome variable are mixed with those of the always-takers and the never-takers. Therefore, the distribution of the subjects assigned to the treatment group and actually taking the treatment, $f_{11}(Y)$, is a mixture of the two distributions: $g_{at}(Y)$, for always-takers, and $g_{c1}(Y)$, for compliers, where

$$f_{11}(y) = \frac{\pi_c}{\pi_{at+c}} g_{c1}(y) + \frac{\pi_{at}}{\pi_{at+c}} g_{at}(y). \tag{25}$$

Similarly, the distribution of the subjects assigned to the control group and those that actually comply with their assignment, $f_{00}(Y)$, is a mixture of the two distributions: $g_{nt}(Y)$, for never-takers and $g_{c0}(Y)$, for compliers, where

$$f_{00}(y) = \frac{\pi_c}{\pi_{nt+c}} g_{c0}(y) + \frac{\pi_{nt}}{\pi_{nt+c}} g_{nt}(y). \tag{26}$$

Equation 25 and 26 can be inverted so that $g_{c1}(Y)$ and $g_{c0}(Y)$ can be expressed with all estimable terms:

$$g_{c1}(y) = \frac{\pi_{at+c}}{\pi_c} f_{11}(y) - \frac{\pi_{at}}{\pi_c} f_{01}(y), \tag{27}$$

and

$$g_{c0}(y) = \frac{\pi_{nt+c}}{\pi_c} f_{00}(y) - \frac{\pi_{nt}}{\pi_c} f_{10}(y). \tag{28}$$

Therefore, the empirical distribution of $g_{c1}(Y)$ and $g_{c0}(Y)$ can be expressed with the observable distributions and sample proportions as

$$\hat{g}_{c1}(y) = \frac{p_{at+c}}{p_c} \hat{f}_{11}(y) - \frac{p_{at}}{p_c} \hat{f}_{01}(y), \tag{29}$$

and

$$\hat{g}_{c0}(y) = \frac{p_{nt+c}}{p_c} \hat{f}_{00}(y) - \frac{p_{nt}}{p_c} \hat{f}_{10}(y). \tag{30}$$

From Equation 29 and Equation 30, both empirical distributions for compliers in the control and the treatment groups can be derived.

The Standard IV method, on the other hand, only derives the means of the distributions. Imbens and Rubin (1997b) further displayed how the Standard IV

method is "implicitly based on" (Imbens & Rubin, 1997b, p. 560) the outcome distribution estimation method. From Equation 27 and Equation 28, it is known that

$$\mu_{g_{c1}(y)} = E\left[Y_i(1) \mid K_i = c\right] = \frac{\pi_{at+c}}{\pi_c} E\left[\overline{Y}_{11}\right] - \frac{\pi_{at}}{\pi_c} E\left[\overline{Y}_{01}\right], \tag{31}$$

and

$$\mu_{g_{c0}(y)} = E\left[Y_i(0) \mid K_i = c\right] = \frac{\pi_{nt+c}}{\pi_c} E\left[\overline{Y}_{00}\right] - \frac{\pi_{nt}}{\pi_c} E\left[\overline{Y}_{10}\right], \tag{32}$$

where $\overline{Y}_{11}$ is the mean of units with $Z = 1$ and $D = 1$, $\overline{Y}_{01}$ with $Z = 0$ and $D = 1$, $\overline{Y}_{00}$ with $Z = 0$ and $D = 0$, and $\overline{Y}_{10}$ with $Z = 1$ and $D = 0$.

If subtracting Equation 32 from Equation 31, the CACE can be calculated as

$$\begin{aligned} CACE = \Delta_c &= \mu_{g_{c1}(y)} - \mu_{g_{c0}(y)} \\ &= \left(\frac{\pi_{at+c}}{\pi_c} E\left[\overline{Y}_{11}\right] + \frac{\pi_{nt}}{\pi_c} E\left[\overline{Y}_{10}\right]\right) - \left(\frac{\pi_{at}}{\pi_c} E\left[\overline{Y}_{01}\right] + \frac{\pi_{nt+c}}{\pi_c} E\left[\overline{Y}_{00}\right]\right) \\ &= \frac{E\left[\overline{Y}_1\right]}{\pi_{c+at} - \pi_{at}} - \frac{E\left[\overline{Y}_0\right]}{\pi_{c+at} - \pi_{at}}. \end{aligned} \tag{33}$$

Equation 33 can also be estimated with Equation 16; therefore, the Standard IV estimation method is essentially based on the outcome distribution estimation method.

The problem with the Standard IV estimation method is that it only uses the expectations of observed Y_i conditional on the observed assignment variable Z_i and the observed treatment taking variable D_i. The underlying mixture structure implied by the model, expressed in Equations 25 and 26, is not taken into account in this approach. Specifically, the observed distributions, $f_{11}(Y)$ and $f_{00}(Y)$, are mixtures of the observable distributions $g_{at}(Y)$ and $g_{nt}(Y)$ and of the non-observable

distributions $g_{c0}(Y)$ and $g_{c1}(Y)$. The density of all these distributions should be constrained to be non-negative. However, while estimating the distributions of interest, $\hat{g}_{c0}(Y)$ and $\hat{g}_{c1}(Y)$, using Equations 29 and 30, negative density is likely to present due to sampling variation or violation of the assumptions.

To demonstrate possible negativity due to sampling variance, simulation data were generated. All assumptions were met to exclude other reasons for negative density. Table 3 presents the means and proportions for subjects with different compliance statuses and different treatment levels received. In specific, compliers assigned to the treatment group had a population mean of 3 and assigned to the control group had a population mean of 0; always-takers had a population mean of 6 no matter which group they were assigned; never-takers had -3 irrespective of the assignment. There were no defiers because of the assumption of monotonicity. Variances were all set at 1, and all distributions were normal. A total sample of 20,000 subjects was selected from all the six distributions and the proportion of each subsample was also specified in Table 3. In order to make our exemplification easy, the proportion of each subsample was set to have no sampling variance—the subsample proportions equaled the subpopulation proportions. Each data point was forced to have only one place after the decimal so that the proportion of each data point approximated the integrated density for a small range of data points nearby. Because of the large sample size, the graph, which plotted the proportion of each data point against each data point, approximated the density plot of each distribution.

Table 3

Subpopulation Means and Proportion

		Actual Treatment Taking (D)	
		1 (Take Treatment)	0 (Not Take Treatment)
Assignmen	1 (Treat)	$\hat{f}_{11}$ Compliers (μ_{c1} = 3, 25%) Always-takers (μ_{c1} = 6, 10%)	$\hat{f}_{10}$ Never-takers (μ_{c1} = -3, 15%)
	0 (Control)	$\hat{f}_{01}$ Always-takers (μ_{c1} = 6, 10%)	$\hat{f}_{00}$ Compliers (μ_{c1} = 0, 25%) Never-takers (μ_{c1} = -3, 15%)

Figure 4 depicts the comparison of the estimated density (symbolled with "●") and the generated density (symbolled with "Δ") for compliers assigned to the treatment group, and Figure 5 presents the comparison for compliers assigned to the control group. The empirical distributions of the two figures were derived using Equations 29 and 30. As it is shown in Figure 4, the generated density is constantly positive, but towards the positive tail, the estimated density has some negative values. A similar pattern can be observed in Figure 5 where the generated density is constantly positive, but towards the negative tail, the estimated density has some negative values.

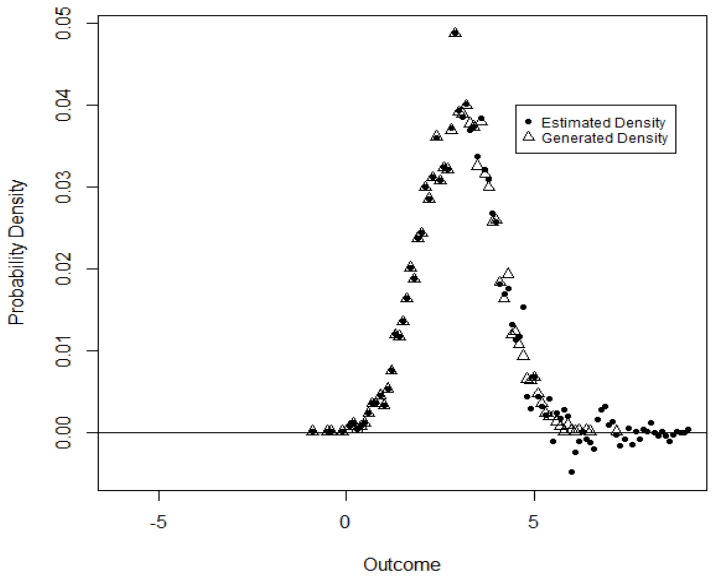

Figure 4. Comparison of the estimated density and the generated density for compliers assigned to the treatment group.

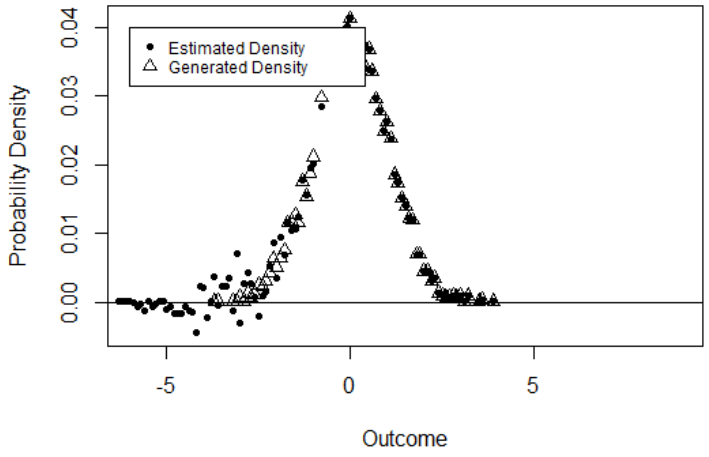

Figure 5. Comparison of the estimated density and the generated density for compliers assigned to the control group.

To enforce non-negativity for $g_{c0}(Y)$ and $g_{c1}(Y)$, Imbens and Rubin (1997b) proposed to build a mixture model and use maximum likelihood to estimate the parameters. They assumed that the outcome variable *Y* for compliers, always-takers, and never-takers all followed a normal distribution, but they have their own variance depending on their own compliance status and their own mean depending on their compliance and treatment level received. The likelihood for observed data $(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$ can be expressed as

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}) \propto & \prod_{i \in \{Z_i=1, D_i=0\}} \pi_{nt} h\left(Y_i \mid \mu_{nt}, \sigma_{nt}^2\right) \times \prod_{i \in \{Z_i=0, D_i=1\}} \pi_{at} h\left(Y_i \mid \mu_{at}, \sigma_{at}^2\right) \\
& \times \prod_{i \in \{Z_i=1, D_i=1\}} \left[\pi_c h\left(Y_i \mid \mu_{c1}, \sigma_{c1}^2\right) + \pi_{at} h\left(Y_i \mid \mu_{at}, \sigma_{at}^2\right)\right] \quad , \qquad (34) \\
& \times \prod_{i \in \{Z_i=0, D_i=0\}} \left[\pi_c h\left(Y_i \mid \mu_{c0}, \sigma_{c0}^2\right) + \pi_{nt} h\left(Y_i \mid \mu_{nt}, \sigma_{nt}^2\right)\right]
\end{aligned}
$$

where $\boldsymbol{\theta} = \boldsymbol{\theta}_{\boldsymbol{k}\in\{c,at,nt\}} = \left(\pi_c, \mu_{c1}, \sigma^2_{c1}, \mu_{c0}, \sigma^2_{c0}, \pi_{at}, \mu_{at}, \sigma^2_{at}, \pi_{nt}, \mu_{nt}, \text{and } \sigma^2_{nt}\right)$ is the set of population parameters for compliers, always-takers, and never-takers, and $h\left(Y \mid \mu, \sigma^2\right)$ represents the normal distribution density function with a mean of μ and a variance of σ^2.

To understand the likelihood above, the joint probability density function is displayed below. The joint probability density function is equal to the likelihood function of the three observed variables, *Y*, *D*, and *Z*, given the set of population parameters $\boldsymbol{\theta}$, $f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta})$,

$$\begin{aligned} f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta}) &= \int f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{K}) f(\boldsymbol{K}) d(\boldsymbol{K}) \\ &= \int f(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{K}) f(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{K}) f(\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{K}) f(\boldsymbol{K}) d(\boldsymbol{K}) \\ &= \sum_{\boldsymbol{k}\in\{c,at,nt\}} f(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}_{\boldsymbol{k}}, \boldsymbol{K}) \Pr(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{K}) \Pr(\boldsymbol{Z}) \Pr(\boldsymbol{K}) \\ &= \prod_{i=1}^{n} \sum_{k_i\in\{c,at,nt\}} f(Y_i \mid Z_i, \boldsymbol{\theta}_{\boldsymbol{k}}, K_i) \Pr(D_i \mid Z_i, K_i) \Pr(Z_i) \Pr(K_i) \end{aligned}, \quad (35)$$

where *n* is the total sample size. Because *Z* is an exogenous variable, $\Pr(\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{K}) = \Pr(\boldsymbol{Z})$. Because *D* can be determined by (*Z*, *K*) and is irrelevant to $\boldsymbol{\theta}$, $f(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{K}) = f(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{K})$ and $\Pr(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{K}) = \Pr(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{K})$. The integration changes into summation because *K* is a categorical variable and hence discrete.

As *D* and *Z* are dichotomous variables, $f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ can be written as

$$
\begin{aligned}
f\left(\boldsymbol{Y},\boldsymbol{D},\boldsymbol{Z}\mid\boldsymbol{\theta}\right) = & \prod_{i\in\{Z_i=1,D_i=0\}} f\left(Y_i,Z_i=1,D_i=0\mid\boldsymbol{\theta}\right) \\
& \times \prod_{i\in\{Z_i=0,D_i=1\}} f\left(Y_i,Z_i=0,D_i=1\mid\boldsymbol{\theta}\right) \\
& \times \prod_{i\in\{Z_i=1,D_i=1\}} f\left(Y_i,Z_i=1,D_i=1\mid\boldsymbol{\theta}\right) \\
& \times \prod_{i\in\{Z_i=0,D_i=0\}} f\left(Y_i,Z_i=0,D_i=0\mid\boldsymbol{\theta}\right)
\end{aligned} \tag{36}
$$

By substituting the conditional form of Equation 35, the joint probability of variables Y, D and Z takes the form

$$
\begin{aligned}
& f\left(\boldsymbol{Y},\boldsymbol{D},\boldsymbol{Z}\mid\boldsymbol{\theta}\right) \\
= & \prod_{i\in\{Z_i=1,D_i=0\}} \sum_{K_i\in\{c,at,nt\}} f\left(Y_i\mid Z_i=1,\boldsymbol{\theta_k},K_i\right)\Pr\left(D_i=0\mid Z_i=1,K_i\right)\Pr(Z_i=1)\Pr(K_i) \\
\times & \prod_{i\in\{Z_i=0,D_i=1\}} \sum_{K_i\in\{c,at,nt\}} f\left(Y_i\mid Z_i=0,\boldsymbol{\theta_k},K_i\right)\Pr\left(D_i=1\mid Z_i=0,K_i\right)\Pr(Z_i=0)\Pr(K_i) \\
\times & \prod_{i\in\{Z_i=1,D_i=1\}} \sum_{K_i\in\{c,at,nt\}} f\left(Y_i\mid Z_i=1,\boldsymbol{\theta_k},K_i\right)\Pr(D_i=1\mid Z_i=1,K_i)\Pr(Z_i=1)\Pr(K_i) \\
\times & \prod_{i\in\{Z_i=0,D_i=0\}} \sum_{K_i\in\{c,at,nt\}} f\left(Y_i\mid Z_i=0,\boldsymbol{\theta_k},K_i\right)\Pr(D_i=0\mid Z_i=0,K_i)\Pr(Z_i=0)\Pr(K_i) \\
= & \prod_{i\in\{Z_i=1,D_i=0\}} [f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_c,K_i=c\right)\Pr(D_i=0\mid Z_i=1,K_i=c)\Pr(Z_i=1)\Pr(K_i=c) \\
& \qquad + f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_{at},K_i=at\right)\Pr(D_i=0\mid Z_i=1,K_i=at)\Pr(Z_i=1)\Pr(K_i=at) \\
& \qquad + f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_{nt},K_i=nt\right)\Pr(D_i=0\mid Z_i=1,K_i=nt)\Pr(Z_i=1)\Pr(K_i=nt)] \\
\times & \prod_{i\in\{Z_i=0,D_i=1\}} [f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_c,K_i=c\right)\Pr(D_i=1\mid Z_i=0,K_i=c)\Pr(Z_i=0)\Pr(K_i=c) \\
& \qquad + f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_{at},K_i=at\right)\Pr(D_i=1\mid Z_i=0,K_i=at)\Pr(Z_i=0)\Pr(K_i=at) \\
& \qquad + f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_{nt},K_i=nt\right)\Pr(D_i=1\mid Z_i=0,K_i=nt)\Pr(Z_i=0)\Pr(K_i=nt)] \\
\times & \prod_{i\in\{Z_i=1,D_i=1\}} [f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_c,K_i=c\right)\Pr(D_i=1\mid Z_i=1,K_i=c)\Pr(Z_i=1)\Pr(K_i=c) \\
& \qquad + f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_{at},K_i=at\right)\Pr(D_i=1\mid Z_i=1,K_i=at)\Pr(Z_i=1)\Pr(K_i=at) \\
& \qquad + f\left(Y_i\mid Z_i=1,\boldsymbol{\theta}_{nt},K_i=nt\right)\Pr(D_i=1\mid Z_i=1,K_i=nt)\Pr(Z_i=1)\Pr(K_i=nt)] \\
\times & \prod_{i\in\{Z_i=0,D_i=0\}} [f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_c,K_i=c\right)\Pr(D_i=0\mid Z_i=0,K_i=c)\Pr(Z_i=0)\Pr(K_i=c) \\
& \qquad + f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_{at},K_i=at\right)\Pr(D_i=0\mid Z_i=0,K_i=at)\Pr(Z_i=0)\Pr(K_i=at) \\
& \qquad + f\left(Y_i\mid Z_i=0,\boldsymbol{\theta}_{nt},K_i=nt\right)\Pr(D_i=0\mid Z_i=0,K_i=nt)\Pr(Z_i=0)\Pr(K_i=nt)].
\end{aligned} \tag{37}
$$

In the above equation, some conditional probabilities are equal to 0 (e.g.,

$\Pr\left(D_i=0\mid Z_i=1,K_i=at\right)=0$, $\Pr\left(D_i=0\mid Z_i=1,K_i=c\right)=0$,

$\Pr\left(D_i=1\mid Z_i=0,K_i=nt\right)=0$, $\Pr\left(D_i=1\mid Z_i=0,K_i=c\right)=0$,

$\Pr\left(D_i=1\mid Z_i=1,K_i=nt\right)=0$, and $\Pr\left(D_i=0\mid Z_i=0,K_i=at\right)=0$)

and some probabilities are equal to 1 (e.g., $\Pr\left(D_i = 1 \mid Z_i = 1, K_i = c\right) = 1$,

$\Pr\left(D_i = 1 \mid Z_i = 0, K_i = at\right) = 1$, $\Pr\left(D_i = 1 \mid Z_i = 1, K_i = at\right) = 1$,

$\Pr\left(D_i = 0 \mid Z_i = 1, K_i = nt\right) = 1$, $\Pr\left(D_i = 0 \mid Z_i = 0, K_i = nt\right) = 1$ and

$\Pr\left(D_i = 0 \mid Z_i = 0, K_i = c\right) = 1$). Therefore, Equation 37 can be further reduced to

$$
\begin{aligned}
& f\left(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta}\right) \\
= & \prod_{i \in \{Z_i=1, D_i=0\}} \left[f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{nt}, K_i = nt\right) \Pr(Z_i = 1) \Pr(K_i = nt)\right] \\
\times & \prod_{i \in \{Z_i=0, D_i=1\}} \left[f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{at}, K_i = at\right) \Pr(Z_i = 0) v(K_i = at)\right] \\
\times & \prod_{i \in \{Z_i=1, D_i=1\}} \left[f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{c}, K_i = c\right) \Pr(Z_i = 1) \Pr(K_i = c)\right. \\
& \quad \left. + f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{at}, K_i = at\right) \Pr(Z_i = 1) \Pr(K_i = at)\right] \\
\times & \prod_{i \in \{Z_i=0, D_i=0\}} \left[f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{c}, K_i = c\right) \Pr(Z_i = 0) \Pr(K_i = c)\right. \\
& \quad \left. + f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{nt}, K_i = nt\right) \Pr(Z_i = 0) \Pr(K_i = nt)\right] \\
= & \Pr(Z_i = 1)^{n_{Z=1}} \times \Pr(Z_i = 0)^{n_{Z=0}} \\
\times & \prod_{i \in \{Z_i=1, D_i=0\}} \left[f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{nt}, K_i = nt\right) \Pr(K_i = nt)\right] \\
\times & \prod_{i \in \{Z_i=0, D_i=1\}} \left[f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{at}, K_i = at\right) \Pr(K_i = at)\right] \\
\times & \prod_{i \in \{Z_i=1, D_i=1\}} \left[f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{c}, K_i = c\right) \Pr(K_i = c) + f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{at}, K_i = at\right) \Pr(K_i = at)\right] \\
\times & \prod_{i \in \{Z_i=0, D_i=0\}} \left[f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{c}, K_i = c\right) \Pr(K_i = c) + f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{nt}, K_i = nt\right) \Pr(K_i = nt)\right]
\end{aligned}
, \qquad (38)
$$

where $f\left(Y_i \mid Z_i, \boldsymbol{\theta}_k, K_i\right)$ follows a normal distribution of $h\left(Y \mid \mu_{KZ}, \sigma_{KZ}\right)$,

$f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_c, K_i = c\right)$ follows $h\left(Y_i \mid \mu_{c1}, \sigma_{c1}\right)$, and $f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_c, K_i = c\right)$

follows $h\left(Y \mid \mu_{c0}, \sigma_{c0}\right)$. As always-takers and never-takers follow the same

distribution irrespective of their treatment assignment, $f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{nt}, K_i = nt\right)$ and

$f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{nt}, K_i = nt\right)$ follow $h\left(Y_i \mid \mu_{nt}, \sigma_{nt}\right)$, and $f\left(Y_i \mid Z_i = 1, \boldsymbol{\theta}_{at}, K_i = at\right)$ and

$f\left(Y_i \mid Z_i = 0, \boldsymbol{\theta}_{at}, K_i = at\right)$ follow $h\left(Y_i \mid \mu_{at}, \sigma_{at}\right)$. In addition, $\Pr\left(Z_i = 1\right)$ and

$\Pr\left(Z_i = 0\right)$ are predetermined by researchers; hence, $\Pr(Z_i = 1)^{n_{Z=1}} \times \Pr(Z_i = 0)^{n_{Z=0}}$ is a

constant. As a result, $f(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z} \mid \boldsymbol{\theta})$ takes the same form as the likelihood function in Equation 34.

Parameters $\boldsymbol{\theta}$ can be estimated with the maximum likelihood using the EM algorithm (Dempster, Laird, & Rubin, 1977; Redner & Walker, 1984). The EM algorithm is an approach that iteratively computes the maximum-likelihood estimates of population parameters that govern the distribution of variables in an observed incomplete data set (Dempster et al., 1977).

In the scenario discussed here, observed values on variables *Y*, *D*, and *Z* along with values on the partially observed class status variable *K* can be viewed as complete data. With only vectors of *Y*, *D*, and *Z*, the data are incomplete and the likelihood function in Equation 34, $L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$, therefore corresponds to incomplete data. If the observed data set is complete, it is possible to analytically find a set of parameters $\hat{\boldsymbol{\theta}}$ that maximize the likelihood function. With the incomplete data, unfortunately, it is impossible to find the analytical expressions for $\boldsymbol{\theta}$.

To incorporate variable *K* into the likelihood function, a matrix of indicator variables is therefore introduced to solve the problem. Let $\boldsymbol{C} = \{\boldsymbol{C}_1', ... \boldsymbol{C}_i', ..., \boldsymbol{C}_n'\}'$. Each $\boldsymbol{C}_i$ is a *k*-dimensional vector of zero-one indicator variables, $\boldsymbol{C}_i = \{c_{i1}, c_{i2}, ..., c_{iK}\}$ and

$$c_{ik} = \begin{cases} 1, & \text{if the } i\text{th subject belongs to class } k \\ 0, & \text{if the } i\text{th subject does not belong to class } k \end{cases}.$$

As a result, the new log-likelihood function, assuming that the newly introduced matrix $\boldsymbol{C}$ can be fully observed, takes the form

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}) &= f\left(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C} \mid \boldsymbol{\theta}\right) \\
&= f\left(\boldsymbol{Y} | \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{Z} \mid \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{C} \mid \boldsymbol{\theta}\right) \\
&= f\left(\boldsymbol{Y} | \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{D} \mid \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{Z}\right) \Pr\left(\boldsymbol{C} \mid \boldsymbol{\theta}\right) \\
&\propto f\left(\boldsymbol{Y} | \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}, \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{C} \mid \boldsymbol{\theta}\right) \\
&= \prod_{i=1}^{n} \prod_{k \in \{c, at, at\}} \left[\pi_k f\left(Y_i \mid D_i, Z_i, \boldsymbol{\theta_k}\right)\right]^{c_{ik}}.
\end{aligned}
\tag{39}
$$

After taking a logarithm on both sides,

$$
\begin{aligned}
\log L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}) &\propto \log \prod_{i=1}^{n} \prod_{k \in \{c, at, at\}} \left[\pi_k f\left(Y_i \mid D_i, Z_i, \boldsymbol{\theta_k}\right)\right]^{c_{ik}} \\
&\propto \sum_{i=1}^{n} \sum_{k \in \{c, at, at\}} c_{ik} \left\{\log\left(\pi_k\right) + \log\left[f\left(Y_i \mid D_i, Z_i, \boldsymbol{\theta_k}\right)\right]\right\},
\end{aligned}
\tag{40}
$$

The EM algorithm manages to solve the problem with two steps for each iteration: the Expectation Step (the E-step) and the Maximization Step (the M-step). The goal of the E-step is to find the expectation of the complete likelihood regarding the unknown data on *C* given the observed data and the parameters estimated from the last iteration $\boldsymbol{\theta}^{(t-1)}$. The expectation function can be defined as

$$
\begin{aligned}
Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right) &= E\left[\log L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{C}) \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\theta}^{(t-1)}\right] \\
&\propto \sum_{i=1}^{n} \sum_{k \in \{c, at, nt\}} E\left[c_{ik} \mid Y_i, D_i, Z_i, \boldsymbol{\theta}^{(t-1)}\right] \left\{\log\left(\pi_k\right) + \log\left[f\left(Y_i \mid D_i, Z_i, \boldsymbol{\theta_k}\right)\right]\right\}.
\end{aligned}
\tag{41}
$$

Note that $E\left[c_{i\_at} \mid Y_i, D_i = 0, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = 0$, $E\left[c_{i\_nt} \mid Y_i, D_i = 1, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = 0$,

$E\left[c_{i\_c} \mid Y_i, D_i = 1, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = 0$, $E\left[c_{i\_at} \mid Y_i, D_i = 0, Z_i = 1, \boldsymbol{\theta}^{(t-1)}\right] = 0$,

$E\left[c_{i\_c} \mid Y_i, D_i = 0, Z_i = 1, \boldsymbol{\theta}^{(t-1)}\right] = 0$, $E\left[c_{i\_nt} \mid Y_i, D_i = 1, Z_i = 1, \boldsymbol{\theta}^{(t-1)}\right] = 0$, and

$E\left[c_{i\_nt} \mid Y_i, D_i = 0, Z_i = 1, \boldsymbol{\theta}^{(t-1)}\right] = 1$, $E\left[c_{i\_at} \mid Y_i, D_i = 1, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = 1$. Therefore,

$$
\begin{aligned}
&Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{(t-1)}\right)\propto \\
&\sum_{i\in\{Z_i=1,D_i=1\}}\left(\begin{array}{l}E\left[c_{i\_c}\mid Y_i,D_i=1,Z_i=1,\boldsymbol{\theta}^{(t-1)}\right]\left\{\log\left(\pi_c\right)+\log\left[f\left(Y_i\mid D_i=1,Z_i=1,\boldsymbol{\theta}_{\boldsymbol{c}}\right)\right]\right\}\\ +E\left[c_{i\_at}\mid Y_i,D_i=1,Z_i=1,\boldsymbol{\theta}^{(t-1)}\right]\left\{\log\left(\pi_{at}\right)+\log\left[f\left(Y_i\mid D_i=1,Z_i=1,\boldsymbol{\theta}_{\boldsymbol{at}}\right)\right]\right\}\end{array}\right)\\
&+\sum_{i\in\{Z_i=0,D_i=0\}}\left(\begin{array}{l}E\left[c_{i\_c}\mid Y_i,D_i=0,Z_i=0,\boldsymbol{\theta}^{(t-1)}\right]\left\{\log\left(\pi_c\right)+\log\left[f\left(Y_i\mid D_i=0,Z_i=0,\boldsymbol{\theta}_{\boldsymbol{c}}\right)\right]\right\}\\ +E\left[c_{i\_nt}\mid Y_i,D_i=0,Z_i=0,\boldsymbol{\theta}^{(t-1)}\right]\left\{\log\left(\pi_{nt}\right)+\log\left[f\left(Y_i\mid D_i=0,Z_i=0,\boldsymbol{\theta}_{\boldsymbol{nt}}\right)\right]\right\}\end{array}\right)\\
&+\sum_{i\in\{Z_i=1,D_i=0\}}\left\{\log\left(\pi_{nt}\right)+\log\left[f\left(Y_i\mid D_i=0,Z_i=1,\boldsymbol{\theta}_{nt}\right)\right]\right\}\\
&+\sum_{i\in\{Z_i=0,D_i=1\}}\left\{\log\left(\pi_{at}\right)+\log\left[f\left(Y_i\mid D_i=1,Z_i=0,\boldsymbol{\theta}_{at}\right)\right]\right\}.
\end{aligned}
\tag{42}
$$

In addition, according to Bayes theorem,

$$
\begin{aligned}
&E\left[c_{ik}\mid Y_i,D_i,Z_i,\boldsymbol{\theta}^{(t-1)}\right]\\
&=\Pr(c_{ik}=1\mid Y_i,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\\
&=\frac{f(Y_i,D_i,Z_i,\mid c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}{f(Y_i,D_i,Z_i\mid\boldsymbol{\theta}^{(t-1)})}\\
&=\frac{f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(D_i,Z_i\mid c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}{\sum_{k\in\{c,at,nt\}}f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(D_i,Z_i\mid c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}\\
&=\frac{f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(D_i\mid Z_i,c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(Z_i\mid c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}{\sum_{k\in\{c,at,nt\}}f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(D_i\mid Z_i,c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(Z_i\mid c_{ik}=1,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}\\
&=\frac{f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}{\sum_{k\in\{c,at,nt\}}f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})}
\end{aligned}
\quad , \tag{43}
$$

whereas, $f(Y_i\mid c_{i\_nt}=1,D_i=1,Z_i=1,\boldsymbol{\theta}^{(t-1)})=0,$

$f(Y_i\mid c_{i\_at}=1,D_i=0,Z_i=0,\boldsymbol{\theta}^{(t-1)})=0,\ \ f(Y_i\mid c_{ik}=1,D_i,Z_i,\boldsymbol{\theta}^{(t-1)})$ follows a normal distribution of $h\left(Y\mid\mu_{KZ}^{(t-1)},\sigma_{KZ}^{2(t-1)}\right)$, and $\Pr(c_{ik}=1\mid\boldsymbol{\theta}^{(t-1)})=\pi_k^{(t-1)}$. As a result,

$$
\begin{aligned}
&E\left[c_{i\_c}\mid Y_i,D_i=1,Z_i=1,\boldsymbol{\theta}^{(t-1)}\right]\\
&=\frac{h(Y_i\mid\mu_{c1}^{(t-1)},\sigma_{c1}^{(t-1)})\pi_c^{(t-1)}}{h(Y_i\mid\mu_{c1}^{(t-1)},\sigma_{c1}^{2(t-1)})\pi_c^{(t-1)}+h(Y_i\mid\mu_{at}^{(t-1)},\sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}},
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&E\left[c_{i\_at}\mid Y_i,D_i=1,Z_i=1,\boldsymbol{\theta}^{(t-1)}\right]\\
&=\frac{h(Y_i\mid\mu_{at}^{(t-1)},\sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}}{h(Y_i\mid\mu_{c1}^{(t-1)},\sigma_{c1}^{2(t-1)})\pi_c^{(t-1)}+h(Y_i\mid\mu_{at}^{(t-1)},\sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}},
\end{aligned}
\tag{45}
$$

$$E\left[c_{i\_c} \mid Y_i, D_i = 0, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = \frac{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}}, \tag{46}$$

and

$$E\left[c_{i\_nt} \mid Y_i, D_i = 0, Z_i = 0, \boldsymbol{\theta}^{(t-1)}\right] = \frac{h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_c^{(t-1)}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}} \tag{47}$$

Therefore,

$$\begin{aligned}
&Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right) \propto \\
&\sum_{i\in\{Z_i=1, D_i=1\}} \left(\frac{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)} \left\{\log(\pi_c) + \log[h(Y_i \mid \mu_{c1}, \sigma_{c1}^2)]\right\}}{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}} \right. \\
&\left. + \frac{h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)} \left\{\log(\pi_{at}) + \log[h(Y_i \mid \mu_{at}, \sigma_{at}^2)]\right\}}{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}} \right) \\
&+ \sum_{i\in\{Z_i=0, D_i=0\}} \left(\frac{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} \left\{\log(\pi_c) + \log[h(Y_i \mid \mu_{c0}, \sigma_{c0}^2)]\right\}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}} \right. \\
&\left. + \frac{h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)} \left\{\log(\pi_{nt}) + \log[h(Y_i \mid \mu_{nt}, \sigma_{nt}^2)]\right\}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}} \right) \\
&+ \sum_{i\in\{Z_i=1, D_i=0\}} \left\{\log(\pi_{nt}) + \log[h(Y_i \mid \mu_{nt}, \sigma_{nt}^2)]\right\} \\
&+ \sum_{i\in\{Z_i=0, D_i=1\}} \left\{\log(\pi_{at}) + \log[h(Y_i \mid \mu_{at}, \sigma_{at}^2)]\right\}.
\end{aligned} \tag{48}$$

In short, the E-step can be viewed as computing the posterior probability of each compliance class with respect to the parameters estimated at the (t-1)-th iteration for each individual in the study.

The M-step then maximizes the expectation function from the E-step with respect to the vector of parameter $\boldsymbol{\theta}$ and yields the parameter estimates for the t-th iteration. By setting the first derivative of $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}\right)$ with respect to each element in

vector $\boldsymbol{\theta}$ equal to 0, the derivative functions can be solved and the updated parameters are obtained as

$$\pi_c^{(t)} = \frac{\sum_{i=1}^{n} E\left[c_{i\_c} \mid Y_i, D_i, Z_i, \boldsymbol{\theta}^{(t-1)}\right]}{n} = \frac{\left(\begin{array}{l} \sum_{i\in\{Z_i=1,D_i=1\}} \frac{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)}}{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}} \\ + \sum_{i\in\{Z_i=0,D_i=0\}} \frac{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}} \end{array}\right)}{n}, \tag{49}$$

$$\pi_{at}^{(t)} = \frac{\sum_{i=1}^{n} E\left[c_{i\_at} \mid Y_i, D_i, Z_i, \boldsymbol{\theta}^{(t-1)}\right]}{n} = \frac{\sum_{i\in\{Z_i=1,D_i=1\}} \frac{h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}}{h(Y_i \mid \mu_{c1}^{(t-1)}, \sigma_{c1}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{at}^{(t-1)}, \sigma_{at}^{2(t-1)})\pi_{at}^{(t-1)}} + n_{Z_i=0,D_i=1}}{n}, \tag{50}$$

$$\pi_{nt}^{(t)} = \frac{\sum_{i=1}^{n} E\left[c_{i\_nt} \mid Y_i, D_i, Z_i, \boldsymbol{\theta}^{(t-1)}\right]}{n} = \frac{\sum_{i\in\{Z_i=0,D_i=0\}} \frac{h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}}{h(Y_i \mid \mu_{c0}^{(t-1)}, \sigma_{c0}^{2(t-1)})\pi_c^{(t-1)} + h(Y_i \mid \mu_{nt}^{(t-1)}, \sigma_{nt}^{2(t-1)})\pi_{nt}^{(t-1)}} + n_{Z_i=1,D_i=0}}{n}, \tag{51}$$

$$\mu_{c1}^{(t)} = \frac{\sum_{i\in\{Z_i=1,D_i=1\}} \left\{E\left[c_{i\_c} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right] \times Y_i\right\}}{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_c} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right]}, \tag{52}$$

$$\mu_{c0}^{(t)} = \frac{\sum_{i\in\{Z_i=0,D_i=0\}} \left\{E\left[c_{i\_c} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right] \times Y_i\right\}}{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_c} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right]}, \tag{53}$$

$$\mu_{at}^{(t)} = \frac{\sum_{i\in\{Z_i=1,D_i=1\}} \left\{E\left[c_{i\_at} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right] \times Y_i\right\} + \sum_{i\in\{Z_i=0,D_i=1\}} Y_i}{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_at} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right] + n_{Z_i=0,D_i=1}}, \tag{54}$$

$$\mu_{nt}^{(t)} = \frac{\sum_{i\in\{Z_i=0,D_i=0\}} \left\{E\left[c_{i\_nt} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right]\times Y_i\right\} + \sum_{i\in\{Z_i=1,D_i=0\}} Y_i}{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_nt} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right] + n_{Z_i=1,D_i=0}}, \tag{55}$$

$$\sigma_{c1}^{2(t)} = \frac{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_c} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right]\left(Y_i-\mu_{c1}^{(t)}\right)^2}{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_c} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right]}, \tag{56}$$

$$\sigma_{c0}^{2(t)} = \frac{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_c} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right]\left(Y_i-\mu_{c0}^{(t)}\right)^2}{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_c} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right]}, \tag{57}$$

$$\sigma_{at}^{2(t)} = \frac{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_at} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right]\left(Y_i-\mu_{at}^{(t)}\right)^2 + \sum_{i\in\{Z_i=0,D_i=1\}} \left(Y_i-\mu_{at}^{(t)}\right)^2}{\sum_{i\in\{Z_i=1,D_i=1\}} E\left[c_{i\_at} \mid Y_i, D_i=1, Z_i=1, \boldsymbol{\theta}^{(t-1)}\right] + n_{Z_i=0,D_i=1}}, \tag{58}$$

and

$$\sigma_{nt}^{2(t)} = \frac{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_nt} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right]\left(Y_i-\mu_{nt}^{(t)}\right)^2 + \sum_{i\in\{Z_i=1,D_i=0\}} \left(Y_i-\mu_{nt}^{(t)}\right)^2}{\sum_{i\in\{Z_i=0,D_i=0\}} E\left[c_{i\_nt} \mid Y_i, D_i=0, Z_i=0, \boldsymbol{\theta}^{(t-1)}\right] + n_{Z_i=1,D_i=0}}. \tag{59}$$

With the newly estimated parameters, the E-step will be repeated again to calculate another set of new posterior probability of each compliance class for every individual and the M-step will continue to yield a new set of parameters. The EM algorithm will eventually stop after the parameters estimated from the M-step reach a negligible change. In the present study, the MMB-EM estimation of CACE built in Mplus 8.1 was used to estimate parameters and their standard errors. Mplus computes the MLR standard errors that are robust to non-normality and non-independence with a sandwich estimator (Muthén & Muthén, 1998-2018).

2.4.Latent Growth Models

As it was discussed in the introduction section, a longitudinal experiment is required when an experimental study focuses on finding long-term effects. In a longitudinal experiment, outcome variables are measured repeatedly. Longitudinal experiments can vary on three main factors: a) randomization occurrence, b) treatment implementation, and c) subjects' compliance status assumption (Gao et al., 2014). Given the current status quo of educational research, the current study only considered baseline treatment randomization and implementation while assuming that subjects' compliance status would not change over time.

In addition, the present study mainly focused on using LGMs to analyze longitudinal experimental data because LGMs are compatible with various data structures, capable of testing hypotheses that focus on individual level changes, and able to improve the precision of the estimated treatment effect with its ability to handle measurement errors at each time point (Hancock et al., 2013).

LGM techniques were first advocated by development specialists (e.g., Bayley, 1956; Bell, 1953, 1954) to investigate individual change or development growth over time. With multiple revisions over time (see, e.g., McArdle, & Epstein, 1987; Rogosa, Randt, & Zimowski, 1982; Rogosa & Willett, 1985), the LGM approach is compatible with various data structures for a wide range of parameter estimations that answer a variety of research questions (Hancock et al., 2013).

LGMs are essentially multilevel models because the repeated measurements are clustered within each person (Little, 2013); therefore, basic multilevel equations

can be applied for LGMs where the Level 1 intercept and slope can be predicted by the Level 2 intercept and slope:

$$\begin{aligned} &\text{Level 1: } y_{iT} = \eta_{\text{Int}\_i}\lambda_{0T} + \eta_{\text{Slp}\_i}\lambda_{1T} + \varepsilon_{iT}, \\ &\text{Level 2: } \eta_{\text{Int}\_i} = \alpha_{\text{Int}} + \zeta_{\text{Int}\_i}, \\ &\qquad\qquad \eta_{\text{Slp}\_i} = \alpha_{\text{Slp}} + \zeta_{\text{Slp}\_i}, \end{aligned} \tag{60}$$

where T is the measurement occasion (T = 1, 2, 3, 4… The maximum number of T depends on the total number of measurements. For convenience, all measurement occasions are assumed to be equally distanced with total measurement occasions of four in this section unless mentioned otherwise.), y_{iT} is the outcome for individual i at measurement occasion T, $\eta_{\text{Int\_i}}$ is the Level 1 intercept for individual i that captures i's initial level performance, $\eta_{\text{Slp\_i}}$ is the Level 1 slope for each individual i as a function of measurement occasion T that captures i's development trajectory, ε_{iT} is the Level 1 error term that follows a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma}^2_{\varepsilon})$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma}^2_{\varepsilon}$ that equals $\begin{bmatrix} \sigma^2_{\varepsilon_1} & & & \\ \sigma_{\varepsilon_1\_\varepsilon_2} & \sigma^2_{\varepsilon_2} & & \\ \cdots & \cdots & \cdots & \\ \sigma_{\varepsilon_1\_\varepsilon_T} & \sigma_{\varepsilon_2\_\varepsilon_T} & \cdots & \sigma^2_{\varepsilon_T} \end{bmatrix}$, α_{Int} and α_{Slp} are the intercept and slope of the Level 2 equations and predict the Level 1 intercept ($\eta_{\text{Int\_i}}$) and slope ($\eta_{\text{Slp\_i}}$), and ζ_{Int_i} and ζ_{Slp_i} are the Level 2 error terms that follow a joint distribution of $N\left(\mathbf{0}, \boldsymbol{\sigma}^2_{\zeta}\right)$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma}^2_{\zeta}$ that equals $\begin{bmatrix} \sigma^2_{\text{Int}} & \\ \sigma_{\text{Int\_Slp}} & \sigma^2_{\text{Slp}} \end{bmatrix}$.

λ_{0T} is fixed to 1 for all Ts (i.e., $\boldsymbol{\lambda_0} = [1,1,1,1]'$). Intuitively speaking, $\eta_{\text{Int}\_i}\lambda_{0T}$ represents participants' initial statuses and their initial statuses do not change over

time, so the loadings from the initial stage should not change. λ_{1T}s are usually prespecified to reflect researchers' belief in the growth trajectory. The most common specification is to fit a linear growth model by fixing, if all measurement occasions are equally distanced, λ_{1T} to be equal to T-1 (i.e., $\boldsymbol{\lambda}_1 = [0,1,2,3]'$) . It is also feasible to specify a nonlinear growth. An unconditional model does not force the growth to be strictly linear; therefore, only $\lambda_{11} = 0$ and $\lambda_{12} = 1$ will be prespecified, and all other $\lambda_{1T(T>2)}$ are freely estimated (Little, 2013). Due to limited time, the current study focused on the more common linear growth models. Further investigation regarding nonlinear models is necessary for a more generalized conclusion. For illustration, Figure 6a depicts a linear LGM with four equally spaced measured time points.

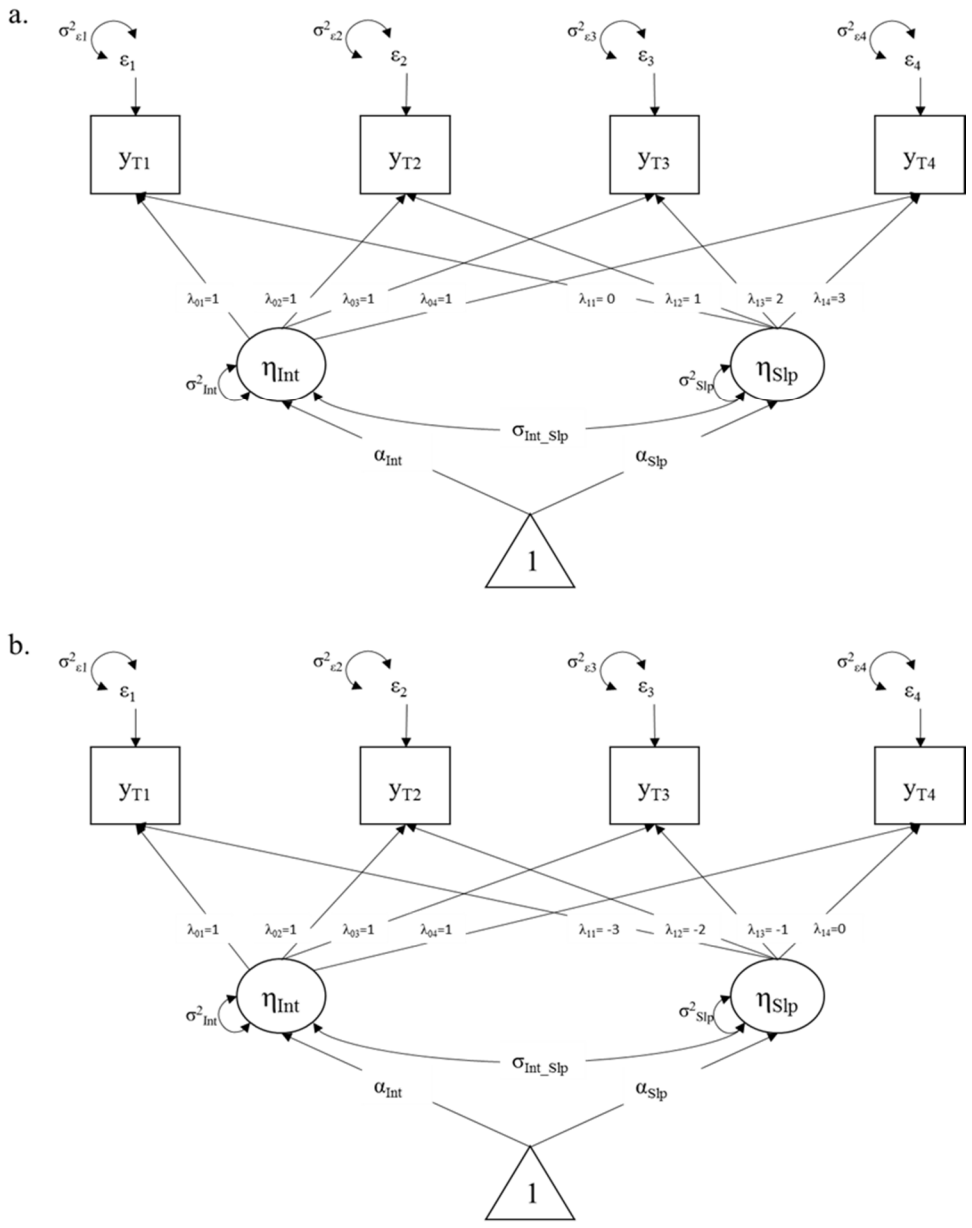

Figure 6. a. Basic form of latent growth models with T1 as the reference point.

b. Basic form of latent growth models with T4 as he reference point.

Within the LGM framework, one can estimate a wide range of parameters and conduct statistical tests for each parameter. In this way, a variety of research questions can be answered (Hancock et al., 2013). First of all, the LGM approach can estimate the means and variances of the latent intercept and the latent slope. The

mean intercept (α_{Int}) quantifies the group level average performance at the beginning, and the mean slope (α_{Slp}) captures average group changing rate over time. Second, the variances of the intercept (σ^2_{Int}) and the slope (σ^2_{Slp}) are also available for estimation. The two parameters respectively describe the variation and the distribution of individuals' initial performances and the variation in their growth trajectories. Third, the LGM approach also allows the estimation of the covariance between the latent intercept and the latent slope ($\sigma_{\text{Int\_Slp}}$). This parameter reflects the relationship between subjects' initial performance and their growth trajectory. Last but not least, by estimating variances of all error terms for each measurement point ($\sigma^2_{e_T}$), the variances that are not a function of the latent growth constituent are eliminated from the observed outcomes.

The error covariances are usually set to 0 (i.e., the off-diagonal elements of the covariance matrix $\boldsymbol{\sigma^2_\varepsilon}$ are all 0) because in most studies theories consider errors as deviations between observed outcomes and expectations from latent models, so they can be results of measurement errors, instrument errors, rater unreliability or model misspecification (Hancock et al., 2013). Therefore, error terms are usually regarded as random at each time point, and error terms at two time points are usually considered uncorrelated. If the theory, however, supports covariance among error terms, one can also add covariances and then conduct a χ^2 difference test to check if the more complicated model fits the data better. In the present study, only error terms with no covariation were considered.

The LGM approach can also adapt to different research assumptions and data structures (Hancock et al., 2013). First, in terms of the reference point, subjects'

reference level can be placed at any time point—the beginning of the study, the end of the study, or even in the middle. In this way, researchers have the freedom to choose a reference point that makes the result interpretation easier. The LGM in Figure 6a uses the baseline time point as the reference level, and Figure 6b presents an LGM that uses the last measurement point as the reference.

Second, LGMs can model both linear and non-linear development trajectories depending on different theoretical hypotheses. The basic linear trend model would only include a single latent linear slope and specify the slope loadings in a way that reflects the linear relationship in accordance with the assessment spacing. With nonlinearity, one can either specify the slope loadings to reflect a suspected nonlinear function for individuals' trajectories (Figure 7a) or even leave the loadings unspecified after the second measurement (Figure 7b).

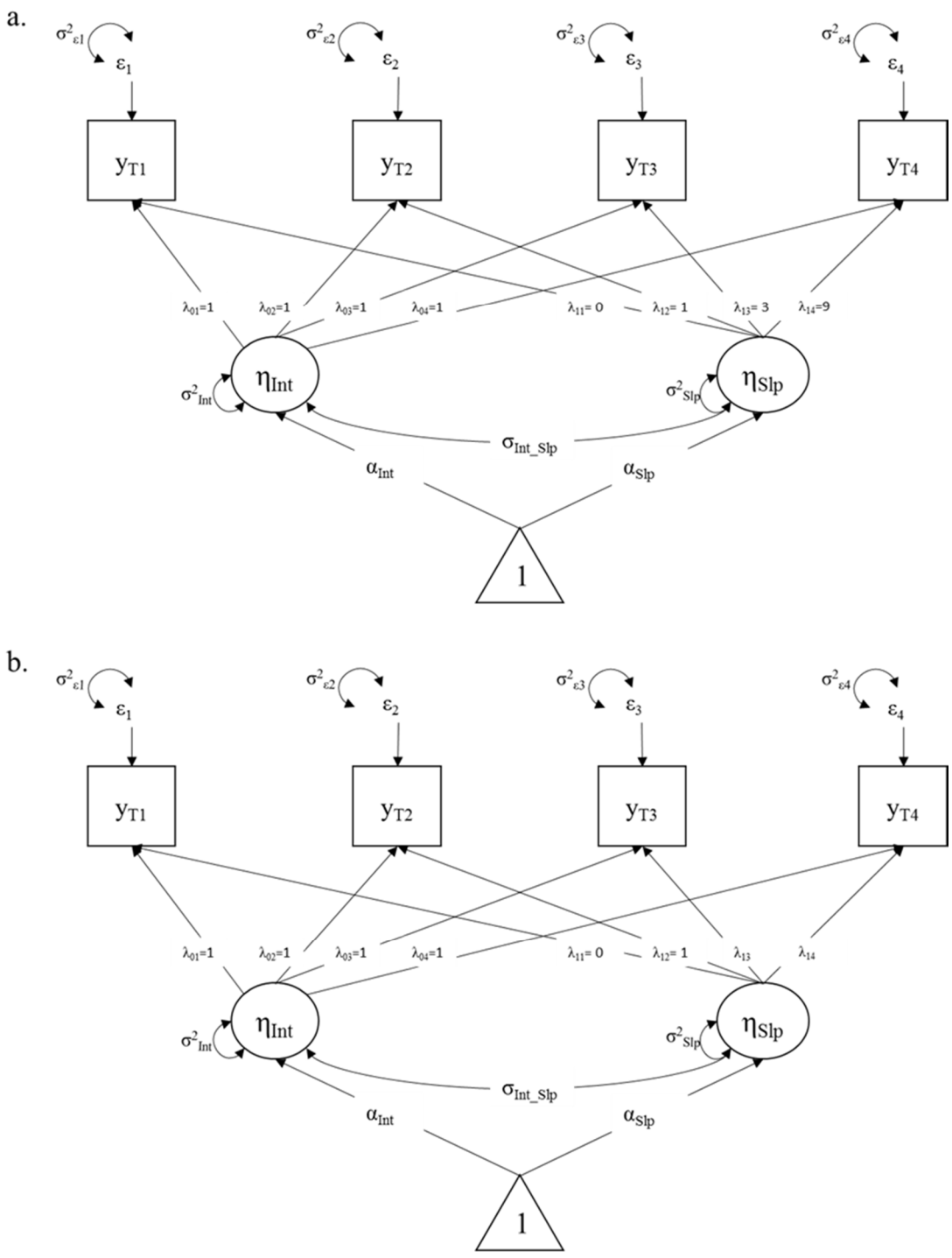

Figure 7. a. Slope loadings reflecting a nonlinear trajectory. b. Slope loadings not completely specified.

If the research interest is to subdivide a set of measurements into stages with theoretical implications and summarize the trend in each stage, piecewise LGMs are extremely useful (Bryk & Raudenbush, 1992). For example, in their study of career role change, Wille, Beyer, and De Fruyt (2012) collected data on seven occasions over 15 years. They placed the reference point at the fourth measurement occasion because it was believed that this point was the turning point where the steeper increase in career roles changed into less noticeable. For both stages, the growth trajectories were linear, but after measurement occasion 4, the linear growth just slowed down. Therefore, they used a piecewise LGM with one latent intercept and two latent slopes (Figure 8a). The first latent slope was associated with T1 through T3 and the second slope was associated with T5 to T7. T4 had a loading of 0 from both slopes because it was the reference point. The two slopes had different means (α_{Slp1} and α_{Slp2}) to reflect different growth rates.

If, instead, multiple functional forms are hypothesized in a study, researchers also have the freedom to include both linear and quadratic factors to model the growth trajectories. Stoolmiller and colleagues (1993) proposed a quadratic growth curve model to analyze maternal resistance during therapy (Figure 8b). In this model, the quadratic growth factor captures the degree of quadratic curvature for all subjects. They fixed the loadings of the quadratic factor to be 1, -2 and 1, with the positive values indicating downward convexity and the negative value indicating upward convexity.

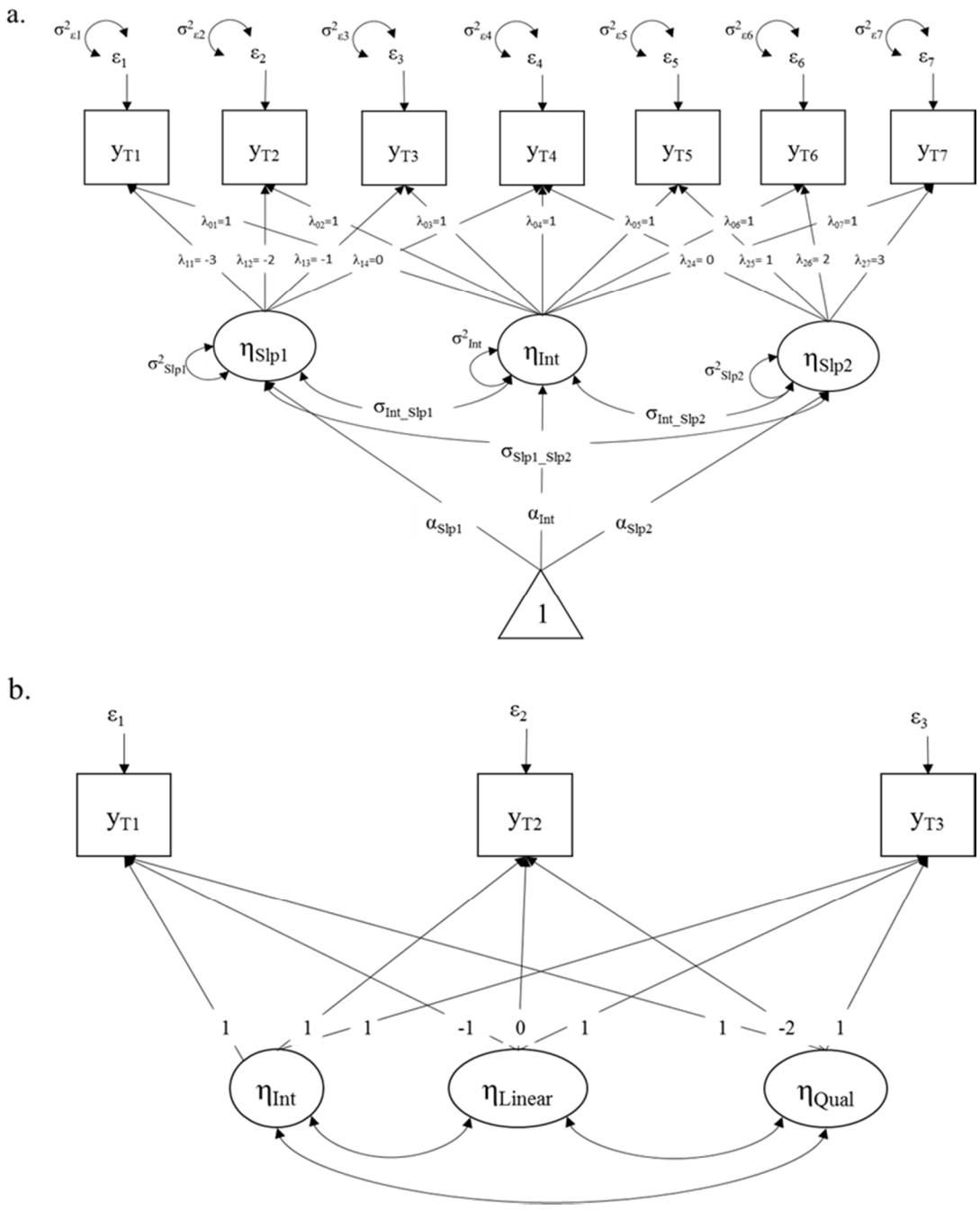

Figure 8. a. Wille, Beyer and De Fruyt's (2012) piecewise growth model for career roles. b. Stoolmiller, Duncan, Bank, and Patterson's (1993) quadratic growth curve model for maternal resistance during therapy.

Last but not least, the LGM approach enables the use of unequally spaced time points. One just needs to specify the coefficients from the latent slope variable(s) for each measurement time point accordingly.

Although the LGM approach has remarkable advantages, it is not a panacea for all longitudinal analyses. Duncan, Duncan, and Strycker (2013) identified its two main limitations. First of all, LGM models, like most SEM models, include the assumption of multivariate normality and the requirement of large samples. The multivariate normality assumption makes it possible to test most parameter statistics. The statistical theory behind LGM is asymptotic in nature that requires a decent sample size to represent the whole population. The second limitation of the LGM approach is that it requires all individuals to be measured for the same amount of times and at the same assessment occasions.

However, simulation studies found that the simplest growth model without missing data nor covariate required a sample size of 40 only in order to have a power of 0.81 to reject the null hypothesis of 0 growth rate (Muthén & Muthén, 2002). The sample size requirement actually depends on a lot of other factors, such as model specification, missing data, and effect size. As for the nonnormality issue, techniques such as bootstrapping (Nevitt & Hancock, 2001) or robust maximum likelihood estimation (Kline, 2012) appear to be promising.

With regard to the second limitation, Duncan et al. (2013) first argued that most longitudinal panel data are typically designed to be collected for a same number of times and at the same occasion for all individuals. They also reasoned that, with certain parameter constraints, even if subjects were not measured for the same

number of times or at the same occasion, it was still possible to apply the LGM. For example, the planned missing data designs (Little & Rhemtulla, 2013) can be applied in LGM when individuals are not all measured at the same occasions. In addition, some software packages (e.g., Mplus) can even accommodate scenarios where individuals are measured at varying time points.

It is important for one to recognize the restrictions of the LGM approach in order to make an educated decision about model selection when analyzing longitudinal data. However, the LGM approach is very powerful and versatile for a wide range of research scenarios. It is flexible even under some limitations. Therefore, it is of great significance to expand the literature of LGM together with noncompliance issues. For simplicity of demonstration, the following section only used the model with a linear trajectory with four equally spaced time points, and there was no covariation among error terms at different measurement occasions. The following discussion can be easily applied to other latent growth models.

2.4.1. Longitudinal experiments with LGMs

As mentioned earlier, the LGM technique can easily adapt to longitudinal experimental scenarios by introducing the treatment variable into the model as the external variable or by using multi-group LGM to estimate the slope differences in the treatment and the control groups.

For example, assume there is a longitudinal experiment with one exogenous categorical treatment assignment variable Z (Z = 1: Treatment; Z = 0: Control), and there is full compliance of treatment assignment (hence treatment effect is the same as treatment assignment effect). If the latent growth trajectories of the two groups are

both assumed to be linear, a multi-group LGM can be used. Within each group, the multilevel equations presented below would hold. The outcome variable *y* for individual *i* assigned to treatment *z* at time point *T* can be expressed as

$$\begin{aligned} &\text{Level 1: } y_{itz} = \eta_{Int\_iz}\lambda_{0T} + \eta_{Slp\_iz}\lambda_{1T} + \varepsilon_{iTz}, \\ &\text{Level 2: } \eta_{Int\_iz} = \alpha_{Int\_z} + \zeta_{Int\_iz}, \\ &\qquad\qquad \eta_{Slp\_iz} = \alpha_{Slp\_z} + \zeta_{Slp\_iz}, \end{aligned} \tag{61}$$

where $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ are the same for both groups ($\boldsymbol{\lambda_0} = [1,1,1,1]'$, $\boldsymbol{\lambda_1} = [0,1,2,3]'$), ζ_{Int_iz} and ζ_{Slp_iz} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}\zeta_z)$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma^2}\zeta_z$ that equals $\begin{bmatrix} \sigma^2_{\text{Int\_z}} & \\ \sigma_{\text{Int\_Slp\_z}} & \sigma^2_{\text{Slp\_z}} \end{bmatrix}$ and ε_{iTz} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2_{\varepsilon\_z}})$ with a mean vector of zeros and covariance matrix $\boldsymbol{\sigma^2_{\varepsilon\_z}}$ that equals $\begin{bmatrix} \sigma^2_{\varepsilon_1\_z} & & & \\ 0 & \sigma^2_{\varepsilon_2\_z} & & \\ 0 & 0 & \sigma^2_{\varepsilon_3\_z} & \\ 0 & 0 & 0 & \sigma^2_{\varepsilon_4\_z} \end{bmatrix}$.

All parameters followed with a subscript *z* indicate that the two treatment levels are essentially two different populations after the experiment. Only $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ are constrained to be the same across the two groups. In fact, by assuming similar patterns and equations for the two groups, the assumption of configural invariance is implied; by holding $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ the same, the weak invariance is assumed (Little, 2013). In the present study, the two invariance assumptions are made because they guarantee that an LGM is a good approximation of the latent growth trajectories for both groups and the two groups are both exhibiting a linear growth pattern. In real data analysis, the configural invariance assumption can be tested easily by fitting the

LGM model separately for each group and check if there is an acceptable model fit (Little, 2013). The weak invariance can be tested with a χ^2 difference test of model fit between a more complicated model with λ_0 and λ_1 differing across the two groups and a more parsimonious model with them being invariant.

The fact that other parameters are different across the treatment and control groups can be viewed as an "effect" of the treatment: by taking the treatment, individuals exhibit higher or low growth rate on average (α_{Slp_z}), the growth rates become more homogeneous or heterogeneous among individuals ($\sigma^2_{\text{Slp}\_z}$), or the ceiling effect becomes more or less prominent ($\sigma_{\text{Int\_Slp\_z}}$).

The parameterization can be simplified for the two latent growth rates. With $Z = 0$, the average slope equals to $\alpha_{\text{Slp}\_z=0}$, and every one-unit increase in *Z* would increase the slope by γ units on average. As the treatment assignment variable *Z* is dichotomous, γ can represent the average slope difference of the group assigned to the treatment level and the group assigned to the control level. Therefore, $\alpha_{\text{Slp}\_z=1} = \alpha_{\text{Slp}\_z=0} + \gamma$ and $\eta_{Slp\_iz} = \alpha_{Slp\_z=0} + \gamma Z_i + \zeta_{Slp\_iz}$.

It is, however, reasonable to argue that other parameters can be the same across the two populations. With the first measurement at baseline before any treatment, the mean intercepts (α_{Int_z}) and the variances of the intercepts ($\sigma^2_{\text{Int\_z}}$) are the same because individuals are randomly assigned to different treatment levels and therefore they have the same expected initial performances and variations. As a result, $\alpha_{\text{Int}\_z=1} = \alpha_{\text{Int}\_z=0} = \alpha_{\text{Int}}$, and $\sigma^2_{\text{Int\_z=1}} = \sigma^2_{\text{Int\_z=0}} = \sigma^2_{\text{Int}}$. In addition, most randomized experiments require that the treatment and control groups differ only on the treatment,

so random errors at a certain measurement occasion are expected to be similar for the two groups: $\boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}\_\mathbf{z=0}} = \boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}\_\mathbf{z=1}} = \boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}$.

These parameters can be estimated, and their equivalence across the treatment and control groups can be tested using the structured mean modeling approach (Thompson & Green, 2013). The structured mean modeling approach is very flexible for scenarios where different groups have different variance structures.

2.4.2. Latent class LGMs

Similarly, if there is no exogenous treatment assignment variable *Z* but only growth mixture, where latent categorical variable *K* has *m* levels of compliance statuses ($k = 1, 2, \ldots, m$), the outcome variable *y* for individual *i* with compliance status *k* at time point *T* can be expressed as

$$\begin{aligned} &\text{Level 1: } y_{itk} = \eta_{\text{Int}\_ik}\lambda_{0T} + \eta_{\text{Slp}\_ik}\lambda_{1T} + \varepsilon_{iTk}, \\ &\text{Level 2: } \eta_{\text{Int}\_ik} = \alpha_{\text{Int}\_k} + \zeta_{\text{Int}\_ik}, \\ &\qquad\qquad \eta_{\text{Slp}\_ik} = \alpha_{\text{Slp}\_k} + \zeta_{\text{Slp}\_ik}, \end{aligned} \tag{62}$$

where $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ are the same for all compliance statuses ($\boldsymbol{\lambda_0} = [1,1,1,1]'$, $\boldsymbol{\lambda_1} = [0,1,2,3]'$), ζ_{Int_ik} and ζ_{Slp_ik} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2\zeta_k})$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma^2\zeta_k}$ that equals $\begin{bmatrix} \sigma^2_{\text{Int}\_k} & \\ \sigma_{\text{Int\_Slp}\_k} & \sigma^2_{\text{Slp}\_k} \end{bmatrix}$ and ε_{iTk} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2_{\varepsilon\_k}})$ with mean vector of zero and covariance matrix $\boldsymbol{\sigma^2_{\varepsilon\_k}}$ that equals $\begin{bmatrix} \sigma^2_{\varepsilon_1\_k} & & & \\ 0 & \sigma^2_{\varepsilon_2\_k} & & \\ 0 & 0 & \sigma^2_{\varepsilon_3\_k} & \\ 0 & 0 & 0 & \sigma^2_{\varepsilon_4\_k} \end{bmatrix}$.

Same as the two-treatment-level model, all parameters with a subscript *k* indicate that each compliance class can have different values on these parameters, and only $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ are constrained to be the same across all compliance classes for the assumptions of configural invariance and weak invariance. If *K* represents compliance statuses, there is no reason to constrain any other parameters to be the same across different compliance classes.

2.4.3. Longitudinal CACE with LGMs

If combining the two parts together—conducting a randomized longitudinal experiment given a population with *m* levels of compliance statuses, the outcome variable *y* for individual *i* with compliance status *k* at time point *T* can be expressed as

$$\begin{aligned} &\text{Level 1: } y_{itzk} = \eta_{\text{Int}\_izk}\lambda_{0T} + \eta_{\text{Slp}\_izk}\lambda_{1T} + \varepsilon_{iTzk}, \\ &\text{Level 2: } \eta_{\text{Int}\_izk} = \alpha_{\text{Int}} + \zeta_{\text{Int}\_izk}, \\ &\qquad\qquad \eta_{\text{Slp}\_izk} = \alpha_{\text{Slp}\_z=0,k} + \gamma_k Z_i + \zeta_{\text{Slp}\_izk}, \end{aligned} \tag{63}$$

where $\boldsymbol{\lambda_0}$ and $\boldsymbol{\lambda_1}$ are the same for all compliance statuses ($\boldsymbol{\lambda_0} = [1,1,1,1]'$, $\boldsymbol{\lambda_1} = [0,1,2,3]'$), ζ_{Int_ikz} and ζ_{Slp_ikz} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2_{\zeta kz}})$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma^2_{\zeta kz}}$ that equals $\begin{bmatrix} \sigma^2_{\text{Int}\_kz} & \\ \sigma_{\text{Int\_Slp}\_kz} & \sigma^2_{\text{Slp}\_kz} \end{bmatrix}$, ε_{iTkz} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2_{\varepsilon\_kz}})$ with mean vector of 0 and covariance matrix $\boldsymbol{\sigma^2_{\varepsilon\_kz}}$ that equals $\begin{bmatrix} \sigma^2_{\varepsilon_1\_kz} & & & \\ 0 & \sigma^2_{\varepsilon_2\_kz} & & \\ 0 & 0 & \sigma^2_{\varepsilon 3\_kz} & \\ 0 & 0 & 0 & \sigma^2_{\varepsilon 4\_kz} \end{bmatrix}$, and γ_k is the effect of the treatment assignment variable *Z* for individuals with compliance status *k* (Jo & Muthén, 2003).

With noncompliance but meeting all assumptions of the IV estimation, there will be three compliance classes—compliers, always-takers, and never-takers. The model in Figure 9 is adapted from Jo and Muthén's (2001) model of growth mixture CACE estimation with repeated outcome measures. *K* represents compliance class. It is surrounded with a circle and a rectangle because compliance classes are partially observable for some individuals. In this model, all parameters can be freely estimated within each compliance status. The arrow from the *K* to the path of *Z* to η_{Int} is set to 0 for always-takers and never-takers, which as a result makes the treatment assignment effect to be 0 for the two subgroups ($\gamma_{at} = \gamma_{nt} = 0$) and only the treatment assignment effect of *Z* for compliers (γ_c) is estimable. Within the complier group, treatment assignment effect equals the treatment effect. More discussion about Figure 9 was included below.

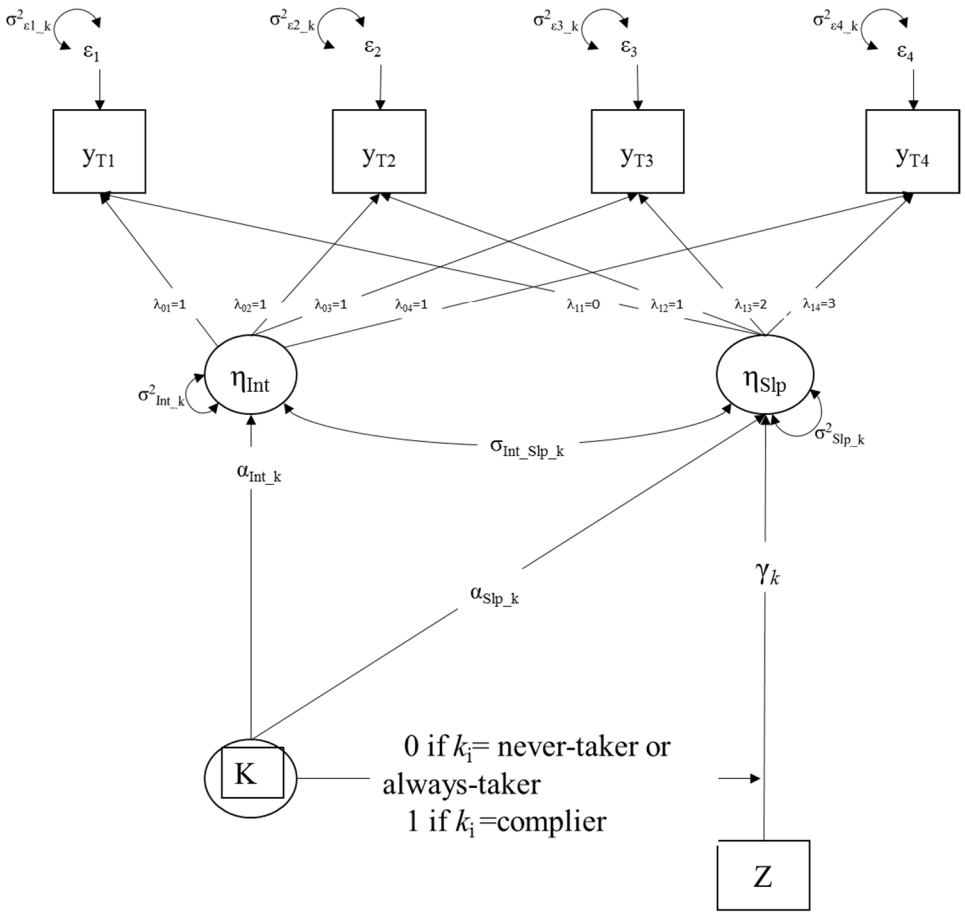

Figure 9. Growth mixture CACE

Although subscript *kz* indicates that individuals of different compliance classes and different treatment levels can have different values on the parameters, some parameters can be constrained to be the same even with different values of *k* or *z*. Other than the parameter constraints discussed earlier, the present study also constrained the Level 2 covariance matrices ($\boldsymbol{\sigma}^2_{\zeta kz}$) to be the same across the treatment and control groups ($\boldsymbol{\sigma}^2_{\zeta kz=0} = \boldsymbol{\sigma}^2_{\zeta kz=1} = \boldsymbol{\sigma}^2_{\zeta k}$). For most intervention studies, the main focuses are the differences in growth trajectories (α_{Slp_z}). Therefore, the mean vectors and covariance matrices for different each compliance classes considered in the present study are the same as included in Table 4.

Table 4

Means and Covariance Matrices for Different Combination of Treatment Levels and Compliance Classes

		Mean vector (μ_{kz})	**Covariance matrix (Σ_{kz})**
Complier	Z=1	$\begin{bmatrix} \alpha_{\text{Int}\_c} \\ \alpha_{\text{Int}\_c}+(\alpha_{\text{Slp}\_c0}+\gamma_c) \\ \alpha_{\text{Int}\_c}+2(\alpha_{\text{Slp}\_c0}+\gamma_c) \\ \alpha_{\text{Int}\_c}+3(\alpha_{\text{Slp}\_c0}+\gamma_c)\lambda_{1T} \end{bmatrix}$	$\begin{bmatrix} \sigma^2_{\text{Int}\_c}+\sigma^2_{e_1\_c} & & & \\ \sigma^2_{\text{Int}\_c}+\sigma_{Int\_Slp\_c} & \sigma^2_{\text{Int}\_c}+2\sigma_{Int\_Slp\_c}+\sigma^2_{\text{Slp}\_c}+\sigma^2_{e2\_c} & & \\ \sigma^2_{\text{Int}\_c}+2\sigma_{Int\_Slp\_c} & \sigma^2_{\text{Int}\_c}+3\sigma_{Int\_Slp\_c}+2\sigma^2_{\text{Slp}\_c} & \sigma^2_{\text{Int}\_c}+4\sigma_{Int\_Slp\_c}+4\sigma^2_{\text{Slp}\_c}+\sigma^2_{e3\_c} & \\ \sigma^2_{\text{Int}\_c}+3\sigma_{Int\_Slp\_c} & \sigma^2_{\text{Int}\_c}+4\sigma_{Int\_Slp\_c}+3\sigma^2_{\text{Slp}\_c} & \sigma^2_{\text{Int}\_c}+5\sigma_{Int\_Slp\_c}+6\sigma^2_{\text{Slp}\_c} & \sigma^2_{\text{Int}\_c}+6\sigma_{Int\_Slp\_c}+9\sigma^2_{\text{Slp}\_c}+\sigma^2_{e4\_c} \end{bmatrix}$
	Z=0	$\begin{bmatrix} \alpha_{\text{Int}\_c} \\ \alpha_{\text{Int}\_c}+\alpha_{\text{Slp}\_c0} \\ \alpha_{\text{Int}\_c}+2\alpha_{\text{Slp}\_c0} \\ \alpha_{\text{Int}\_c}+3\alpha_{\text{Slp}\_c0} \end{bmatrix}$	
Always-taker		$\begin{bmatrix} \alpha_{\text{Int}\_a} \\ \alpha_{\text{Int}\_a}+\alpha_{\text{Slp}\_a} \\ \alpha_{\text{Int}\_a}+2\alpha_{\text{Slp}\_a} \\ \alpha_{\text{Int}\_a}+3\alpha_{\text{Slp}\_a} \end{bmatrix}$	$\begin{bmatrix} \sigma^2_{\text{Int}\_at}+\sigma^2_{e_1\_at} & & & \\ \sigma^2_{\text{Int}\_at}+\sigma_{Int\_Slp\_at} & \sigma^2_{\text{Int}\_at}+2\sigma_{Int\_Slp\_at}+\sigma^2_{\text{Slp}\_at}+\sigma^2_{e2\_at} & & \\ \sigma^2_{\text{Int}\_at}+2\sigma_{Int\_Slp\_at} & \sigma^2_{\text{Int}\_at}+3\sigma_{Int\_Slp\_at}+2\sigma^2_{\text{Slp}\_at} & \sigma^2_{\text{Int}\_at}+4\sigma_{Int\_Slp\_at}+4\sigma^2_{\text{Slp}\_at}+\sigma^2_{e3\_at} & \\ \sigma^2_{\text{Int}\_at}+3\sigma_{Int\_Slp\_at} & \sigma^2_{\text{Int}\_at}+4\sigma_{Int\_Slp\_at}+3\sigma^2_{\text{Slp}\_at} & \sigma^2_{\text{Int}\_at}+5\sigma_{Int\_Slp\_at}+6\sigma^2_{\text{Slp}\_at} & \sigma^2_{\text{Int}\_at}+6\sigma_{Int\_Slp\_at}+9\sigma^2_{\text{Slp}\_at}+\sigma^2_{e4\_at} \end{bmatrix}$
Never-taker		$\begin{bmatrix} \alpha_{\text{Int}\_n} \\ \alpha_{\text{Int}\_n}+\alpha_{\text{Slp}\_n} \\ \alpha_{\text{Int}\_n}+2\alpha_{\text{Slp}\_n} \\ \alpha_{\text{Int}\_n}+3\alpha_{\text{Slp}\_n} \end{bmatrix}$	$\begin{bmatrix} \sigma^2_{\text{Int}\_nt}+\sigma^2_{e_1\_nt} & & & \\ \sigma^2_{\text{Int}\_nt}+\sigma_{Int\_Slp\_nt} & \sigma^2_{\text{Int}\_nt}+2\sigma_{Int\_Slp\_nt}+\sigma^2_{\text{Slp}\_nt}+\sigma^2_{e2\_nt} & & \\ \sigma^2_{\text{Int}\_nt}+2\sigma_{Int\_Slp\_nt} & \sigma^2_{\text{Int}\_nt}+3\sigma_{Int\_Slp\_nt}+2\sigma^2_{\text{Slp}\_nt} & \sigma^2_{\text{Int}\_nt}+4\sigma_{Int\_Slp\_nt}+4\sigma^2_{\text{Slp}\_nt}+\sigma^2_{e3\_nt} & \\ \sigma^2_{\text{Int}\_nt}+3\sigma_{Int\_Slp\_nt} & \sigma^2_{\text{Int}\_nt}+4\sigma_{Int\_Slp\_nt}+3\sigma^2_{\text{Slp}\_nt} & \sigma^2_{\text{Int}\_nt}+5\sigma_{Int\_Slp\_nt}+6\sigma^2_{\text{Slp}\_nt} & \sigma^2_{\text{Int}\_nt}+6\sigma_{Int\_Slp\_nt}+9\sigma^2_{\text{Slp}\_nt}+\sigma^2_{e4\_nt} \end{bmatrix}$

As it is shown in Table 4, there are two separate mean vectors for compliers assigned to the treatment and control groups and the treatment effect is represented with γ_c. Because of the constraints specified earlier ($\boldsymbol{\sigma}^2_{\zeta kz=0} = \boldsymbol{\sigma}^2_{\zeta kz=1} = \boldsymbol{\sigma}^2_{\zeta k}$ and $\boldsymbol{\sigma}^2_{\varepsilon k\_z=0}=\boldsymbol{\sigma}^2_{\varepsilon k\_z=1}=\boldsymbol{\sigma}^2_{\varepsilon k}$), compliers assigned to the treatment group have the same covariance matrix $\boldsymbol{\Sigma}_c$ ($\boldsymbol{\Sigma}_{kz} = \Lambda\boldsymbol{\sigma}^2{}_{\zeta kz}\Lambda' + \boldsymbol{\sigma}^2{}_{\varepsilon\_kz}$, $\Lambda = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$). There are no separate rows for always-takers who are assigned to the treatment group and those who are assigned to the control group for the mean vector and the covariance matrix. This is because always-takers will end up in the treatment group irrespective of the assignment, so there is no difference between the two treatment assignment groups. A similar explanation applies to the never-takers.

The modeling is reflected in Figure 9 too. Compliance class *K* has arrows pointing to the intercept factor and the growth factor, meaning that each compliance class has its own initial status and growth rate. Treatment assignment Z has only one arrow pointing to the slope factor, indicating that the treatment assignment adds an additional part to the growth rate. The additional part, however, becomes 0 for never-takers and always-takers because of the arrow from *K* to the path of *Z* to η_{Slp}. All parameters followed with subscript *k* can be estimated separately for that compliance class.

The likelihood of the observed data $(\boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z})$ ($\boldsymbol{Y} = \{\boldsymbol{Y}_{T1}, \boldsymbol{Y}_{T2}, \boldsymbol{Y}_{T3}, \boldsymbol{Y}_{T4}\}$), adapted from Yau and Little's (2001) work, can be expressed with the following form:

$$
\begin{aligned}
L(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{D}, \boldsymbol{Z}) \propto & \prod_{i \in \{Z_i=1, D_i=0\}} \pi_{nt} h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{nt}, \boldsymbol{\Sigma}_{nt}) \times \prod_{i \in \{Z_i=0, D_i=1\}} \pi_{at} h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{at}, \boldsymbol{\Sigma}_{at}) \\
& \times \prod_{i \in \{Z_i=1, D_i=1\}} \left[\pi_c h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{c1}, \boldsymbol{\Sigma}_c) + \pi_{at} h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{at}, \boldsymbol{\Sigma}_{at}) \right] \\
& \times \prod_{i \in \{Z_i=0, D_i=0\}} \left[\pi_c h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{c0}, \boldsymbol{\Sigma}_c) + \pi_{nt} h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{nt}, \boldsymbol{\Sigma}_{nt}) \right],
\end{aligned} \tag{64}
$$

where $h(\boldsymbol{Y} \mid \boldsymbol{\mu}_{kz}, \boldsymbol{\Sigma}_k)$ follows a multivariate normal distribution with a mean vector $\boldsymbol{\mu}_{kz}$ and a covariance matrix $\boldsymbol{\Sigma}_k$.

The MMB based method can be used to estimate all parameters included in Table 4. Parametric standard errors can be computed with the observed information matrix of the maximum likelihood estimator. The MMB-EM estimation of CACE built in Mplus 8.1 was used to estimate parameters and their standard errors.

2.4.4. Standard IV with LGMs

Standard IV estimation approach will also yield an asymptotically unbiased estimation of γ_c when all assumptions are met. Figure 10 is adapted from the Standard IV estimation model (Figure 2) with a single outcome variable. Similar to the MMB estimation approach, the intercept factor is not affected by the treatment. Exogenous treatment variable *Z* has an effect on the slope factor through assignment-taken variable *D*. The error terms of *D* and the slope factor are allowed to covary, which separates the variation that is not "caused" by *Z* in *D* and the slope factor from the variation that is "caused" by *Z*. In this way, only the exogenous part in *D* is used to determine its effect on *Y*, which equals to γ.

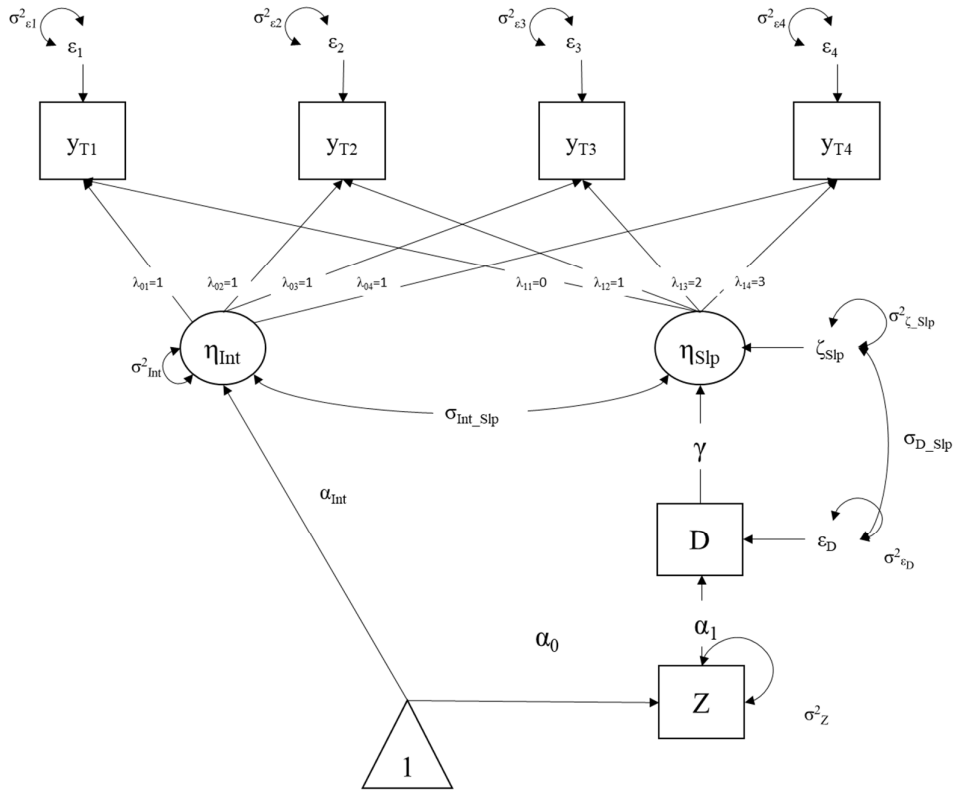

Figure 10. Longitudinal CACE with Standard IV estimation

With the model specified in Figure 10, the outcome variable *y* for the individual *i* at time point *T* can be expressed as

$$\begin{aligned}
&\text{Level 1: } y_{it} = \eta_{\text{Int}_i}\lambda_{0T} + \eta_{\text{Slp}_i}\lambda_{1T} + \varepsilon_{iT}, \\
&\text{Level 2: } \eta_{\text{Int}_i} = \alpha_{\text{Int}} + \zeta_{\text{Int}_i}, \\
&\qquad\quad \eta_{\text{Slp}_i} = \alpha_{\text{Slp}} + \gamma D_i + \zeta_{\text{Slp}_i}, \\
&\text{Level 3: } D_i = \alpha_D + \alpha_1 Z_i + \varepsilon_{D_i}, \\
&\text{Level 4: } Z_i = \alpha_0 + \varepsilon_{Z_i},
\end{aligned} \tag{65}$$

where $\boldsymbol{\lambda_0} = [1,1,1,1]'$, $\boldsymbol{\lambda_1} = [0,1,2,3]'$, $\varepsilon_Z, \varepsilon_D, \zeta_{Int}$ and ζ_{Slp} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}_{\zeta k\zeta})$with mean vector of 0 and covariance matrix $\boldsymbol{\sigma^2}_{\zeta k\zeta}$ that equals

$$\begin{bmatrix} \sigma_Z^2 & & & \\ 0 & \sigma_{\varepsilon_D}^2 & & \\ 0 & \sigma_{D\_Slp} & \sigma_{\zeta_{Slp}}^2 & \\ 0 & 0 & \sigma_{Int\_Slp} & \sigma_{\zeta_{Int}}^2 \end{bmatrix}$$

, ε_{iT} follow a joint distribution of $\boldsymbol{N}(\mathbf{0}, \boldsymbol{\sigma^2}_{\varepsilon})$ with mean

vector of 0 and covariance matrix $\boldsymbol{\sigma}^2_{\boldsymbol{\varepsilon}}$ that equals $\begin{bmatrix} \sigma^2_{\varepsilon_1} & & & \\ 0 & \sigma^2_{\varepsilon_2} & & \\ 0 & 0 & \sigma^2_{\varepsilon 3} & \\ 0 & 0 & 0 & \sigma^2_{\varepsilon 4} \end{bmatrix}$, and γ is the effect of the local treatment effect.

As mentioned in the previous section, with a binary assignment-take variable *D*, the standard errors estimated would be inaccurate because conventional SEM requires continuous dependent variables (Edwards et al., 2012) that are multivariate normal and have homoscedastic residuals (Kline, 2012). In addition, covariance matrices $\boldsymbol{\Sigma}_k$ can be heterogeneous across compliance classes, which also lead to inaccurate standard errors estimated. Fortunately, bootstrapped standard errors are available for an empirical estimation of all standard errors. Therefore, a bootstrapped standard error was used for the Standard IV approach.

2.4.5. Previous studies on the Standard IV approach and the MMB approach

Little and Yau (1998) first used the MMB method to estimate the CACE in a longitudinal study, but they did not analyze the longitudinal data within the framework of LGM. Nevertheless, the MMB approach was quickly included by other researchers to the LGM families (Jo & Muthén, 2001; Jo & Muthén, 2003; Muthén & Asparouhov, 2008). The Standard IV approach, on the other hand, was widely used in both cross-sectional studies (e.g., Altonji, Elder, & Taber, 2005; Brookhart, Wang, Solomon, & Schneeweiss, 2006; Kang & Sivaramakrishnan, 1995) and longitudinal studies (e.g., Bolton & Drew, 1991; Glewwe, Jacoby, & King, 2001; Hogan & Lancaster, 2004). However, the combination of the LGM and the Standard IV approach was not used in most previous longitudinal studies. One possible reason is that the Standard IV approach was mostly used in economic studies where latent

variables were rarely used. Therefore, the author did not find any literature that compared the two estimation methods for longitudinal CACE estimations within the LGM framework.

The two estimation methods were compared within the cross-sectional settings. Imbens and Rubin (1997b) laid out the theoretical foundation for the MMB approach. They argued that the Standard IV method was essentially estimating the first moments of the two marginal distributions of the outcome variable under different treatment assignments for compliers but the MMB approach estimated the whole distribution for compliers first and then obtained the means from each distribution. The former was asymptotically equivalent to the latter, but with small samples or violations of the assumptions, results from the two estimation methods could deviate considerably.

The authors first compared the estimation results of the two approaches using real data. The data was from the Angrist and Krueger's (1991) study of the causal effect of education on earnings. Imbens and Rubin redefined the treatment variable as dichotomous (i.e., with or without twelve or more years of education) and the instrument was subjects' quarter of birth (i.e., born in the first or the fourth quarter). The outcome variable was their weekly earning. The rough estimation of complier percentage was about 2%. They calculated the treatment effect with three estimation approaches: the OLS regression method (ITT), the Standard IV method, and three distributional based methods, including a normal MMB method (using the EM algorithm), a nonnegative IV method (using histogram estimates), and a multinomial MMB method (using constant density within a bin). Consistent with previous studies, the OLS method yielded the smallest treatment effect. However, the Standard IV

method, which was supposed to yield an asymptotically equivalent treatment effect as the three distribution based methods, overestimated the treatment effect.

To interpret the findings, they conducted a small simulation study. They chose a sample size of 1,000 with $\Pr(Z = 1) = 0.5$. There were 10% compliers and 45% always-takers and never-takers. Compliers with $Z = 1$ had a mean of 0.5, and the other half of compliers had a mean of −0.5. Always-takers and never-takers both have a mean of 0. Therefore, the CACE was 1, and the ITT was 0.1. All distributions had a unit variance. The same set of methods mentioned above was used in the simulation study as well. Among all five methods, the normal MMB method had the smallest bias. The nonnegative IV and multinomial MMB methods were more biased than both the normal MMB and Standard IV methods, but the MMB methods had much smaller parameter variations (more efficient) than the Standard IV method. All three distributional based methods tended to yield smaller effects than the real effect. The result indicated that the normal MMB method was the least biased and the most efficient among all methods.

In terms of estimating the first moments for different distributions, both the real data analysis and the simulation study suggested that the normal MMB method was the best in terms of estimation accuracy and efficiency. For the higher order moments, the real data analysis results suggested that when the variances were not properly restricted, the estimation would be highly unreliable with the weak instrument (2% complier percentage), which agreed with common mixture model estimation results. As all distributions were set to have a unit variance, the simulation study did not include the estimation of variances for different distributions.

The aim of Imbens' and Rubin's (1997b) paper was to lay a theoretical background for the MMB method, so their simulation study did not include more

conditions to compare this approach to the Standard IV method. Recognizing its "potential for increased efficiency of estimation over the traditional instrumental variable approach" (p. 157), Little and Yau (1998) called for more simulation studies to compare these two approaches.

Jo (2002) conducted an extensive simulation study to compare how the empirical power in the ITT analysis and the MMB approach for CACE estimation were affected by noncompliance. The impact of four main factors was investigated: compliance rate, study design on cost control, outcome distribution, and covariate information. Only findings regarding the impact of compliance rate and outcome distribution on the MMB method were reviewed here. All results were based on 1,000 replications. Power was defined as the proportion of significant results at α level of 0.05. $\Pr(Z = 1) = 0.5$ was used for all conditions. Only compliers and never-takers were involved, which was consistent with both the JOBS II intervention study (Vinokur et al., 1995) and the JHU PIRC study (Ialongo et al., 1999).

When looking into the effect of compliance rate, three rates were included: 30%, 50%, and 70%. Compliers and never-takers both had a mean of 1.5, but the effect size for compliers were set to be 0.2, 0.5, and 0.8 across all conditions. All designs had a residual variance of 1.0 for both compliers and never-takers. The results suggested that compliance rate had a critical influence on maintaining statistical power. Therefore, with noncompliance, in order to have enough power, researchers should plan a bigger sample size, which usually led to a higher cost.

When investigating the outcome distribution factor, the author manipulated means and outcome variances to investigate their influence on power. In both designs, the sample size was 300, compliers had an effect size of 0.5, and the compliance rate was fixed at 50%. First, when means were manipulated, the outcome was 1.0 for both

compliers and never-takers. The group mean was 1.0 for compliers but ranged from 0.0 to 3.0 for never-takers. The results showed that the CACE analysis benefitted from more separated means: the larger the mean difference, the higher the statistical power. Second, when outcome variance was manipulated, both compliers and never-takers had a mean of 1.5. The outcome variance was 1.0 for compliers and ranged from 0.25 to 2.0 for never-takers. As a result, for the CACE analysis, when the variance of never-takers became smaller than the variance of compliers (set to 1.0), more heterogeneity between these two variances led to higher statistical power. If the variance of never-takers became bigger than the variance of compliers, the statistical power remained stable. The author argued that the reason was that the CACE estimation would benefit from more distinguishable compliers and never-takers, but as the variance of never-takers became too large, the power in estimating the mean of never-takers would be compromised, which directly led to a less efficient estimation of the complier means.

This study explored several important factors that would influence statistical power for the CACE analysis using the MMB method, and for each factor, a wide range of levels were also considered to fully investigate the factor. The results provided very useful information for practitioners to obtain more statistical power, such as planning a bigger sample size accordingly based on different compliance rate.

However, this study did not include the Standard IV method as a comparison, so there was no information about the statistical power when using the Standard IV method. In addition, only two compliance groups were considered, and there was no evidence indicating that similar results could be extended to research scenarios where compliers, never-takers, and always-takers were all involved.

Compliance rate has been shown to be a major concern in previous studies (Imbens & Rubin, 1997b; Little & Yau, 1998). However, in this study, only in the first part (compliance rate and power), the compliance rate was investigated with three different values. For all other factors, the compliance rate was set to 50%. It would be useful to include compliance rate in the investigation of the outcome distribution factor and check if there was an interaction between the two factors. More low compliance rates could have been included because low compliance rate or weak instrument had been suggested to be a problem for instrumental variable analyses (Imbens & Rubin, 1997b; Rothenberg, 1984).

Jo and Muthén (2003) expanded the investigation to longitudinal models. This study tried to use two models to estimate the treatment effect. The first model was an LGM model that defined the treatment effect as the difference in growth trajectories with and without treatment. The second model was an ANOVA model that defined the treatment as the difference in outcome measure at the last time point while controlling for the first time measurement. For all conditions, 500 replications were used with a sample size of 300 and $\Pr(Z = 1) = 0.5$ 50%. They chose the true parameter values, effect size, and sample size, based on the LGM estimation result of the JHU PIRC study. The outcome variable was set to be measured four times with equal distances, and a linear trajectory was used. One thing worth pointing out is that the process of data generation did not strictly follow all assumptions of instrumental variable analysis (Angrist et al., 1996). In this study, two subpopulations were involved: high compliers and low compliers. High compliers had a treatment effect of 0.2 and low compliers had a treatment effect of −0.1. Therefore, there was one more parameter—treatment effect for low compliers—included in the estimation, whereas in the MMB estimation for CACE from Imbens and Rubin (1997b), the treatment

effects for both always-takers and never-takers were 0. This condition was essentially a violation of the exclusion restriction, which cannot be handled with the Standard IV method, and the estimation using the MMB approach was, in fact, a full mixture modeling that estimated the low or high compliance membership for all individuals. The authors used estimation coverage and empirical power rate as two evaluation indices. Coverage was defined as the proportion of replications (out of 500) that had their 95% confidence intervals of treatment effects for both high and low compliers covering the true treatment effects. Empirical power was defined as the percentage of replications (out of 500) that had both treatment effects being significant at the 0.05 level.

Simulation results showed that the point estimates of the treatment effects and their standard errors were similar for both the LGM and the ANOVA models. Both models yielded a high coverage, more than 91% for both high and low compliers, but the power rates were much higher for the LGM model. The authors explained that the growth model used information from all four time points and excluded measurement error from the model; therefore, the power was higher.

Due to a different research interest, the author did not probe deeper on how different factors would influence the coverage or power, but this study demonstrated that the LGM had great potential in estimating longitudinal CACE, which justified its usage in longitudinal experiments.

Very limited research has been done to investigate the CACE estimation within the LGM framework. However, Tolvanen (2007) conducted an intensive simulation study investigating the functionality of the mixture LGM with respect to various factors. Specifically, measurement reliability, separation among latent classes, and sample size exemplified positive influence on estimation convergence rate,

parameter estimation bias, and standard error estimation bias. As the CACE estimation is partially mixture modeling, it is expected to have similar results. However, to what extent would they influence the estimation was still unclear.

In summary, previous studies have recognized the great potential of the MMB approach in estimating the CACE under the LGM framework. However, very little was known about factors that might influence the accuracy and efficiency of longitudinal CACE estimation. Especially with the three compliance statuses condition, only one simulation study (Imbens & Rubin, 1997b) actually included all three statuses. In addition, the Standard IV was not involved in most simulation studies despite its great popularity among researchers.

2.5.Objective of the Present Study

This study purported to investigate the performance of two widely used longitudinal CACE estimation approaches, the Standard IV approach and the MMB approach, under different research scenarios. Past studies shed some light on this question, but investigations on the MMB approach mainly focused on cross-sectional designs (Jo, 2002). However, there was no evidence regarding how the same conclusions could be generalized to longitudinal CACE estimation.

In addition, previous studies used only two compliance classes to generate data. It was unclear how the identified factors would affect the CACE estimation with one more compliance class. In fact, many real data applications included compliers, always-takers, and never-takers (e.g., Angrist & Krueger, 1991; Brookhart et al., 2006). The factors identified to affect the CACE estimation could have a different influence when one more compliance class was included.

Moreover, the Standard IV method would yield an asymptotically unbiased estimation of the CACE when all assumptions were met. However, with a weak instrument or small sample size, the Standard IV approach could be less accurate and less efficient than the MMB approach (Imbens & Rubin, 1997b). However, LGMs could improve the precision in compliance status estimation by using the trajectory information (Jo & Muthén, 2003), but it was unclear whether this feature of LGM techniques would narrow the gap in performance between the Standard IV and the MMB approaches.

Last but not least, past studies highlighted the influence of some factors, such as compliance rate, separation among latent classes, and variance difference among different compliance classes, but the influence of compliance rate was not investigated enough, especially for its interaction effect with other factors. Another important factor, measurement reliability, which has been demonstrated as a crucial factor for mixture modeling was not investigated in the CACE literature. As a result, further investing the complier rate factor and including the measurement reliability factor was another important goal of the present study.

To address the issues that were not covered in previous studies, the present study tried to simulate scenarios where noncompliance was a problem for longitudinal experiments. Particularly, three compliance statuses were simulated, including never-compliers, always-takers, and compliers. In order to cover as much factors that may influence a research design, six factors were considered: sample size, effect size, reliability of measurements, compliance composition, distances among mean latent intercepts and mean latent slopes of the three compliance classes (referred as "mean distance" below), and differences among variances of latent intercepts and slopes of the three compliance classes (referred as "noncomplier-complier Level 2 covariance

ratio" below). For the purpose of estimation methods comparison, both the Standard IV and the MMB based approaches were applied to each simulated dataset to estimate the longitudinal treatment effect respectively. In the end, results regarding estimation success rate, estimation bias, statistical power, and type I error rate were analyzed with respect to the six simulation factors and two estimation approaches. More detailed study design is included in Chapter 3.

.

Chapter 3: Method

Previous chapters discuss that estimating the longitudinal CACE within the framework of the LGM is a solution for longitudinal experiments where noncompliance is an issue. On one hand, the LGM framework has a wide range of applications because of its ability to compare various data structures, capability of testing hypotheses that focus on individual-level changes, and ability to improve the precision of the estimated treatment effect by eliminating measurement errors at each time point and utilizing information from all measurement points. On the other hand, the CACE estimation is able to provide an estimation of the average causal effect for the compliers by using either the Standard IV approach or the MMB approach. Both approaches can easily adapt to the LGM framework for longitudinal CACE estimation. However, there is a need to know how the two approaches would perform within the LGM framework with respect to different research scenarios. The present study aimed to answer five research questions using a simulation design.

This study assumed that an experiment was conducted by randomly selecting a sample from a population and randomly assigning each subject to the treatment group or the control group. For each sample subject, there was a 50% chance that he or she was assigned to the treatment group ($Z = 1$) or the control group ($Z = 0$). The population had three underlying subpopulations with different compliance behaviors. Subjects' compliance behaviors were only partially observable, so a latent compliance membership variable *K* was used to describe subjects' compliance statuses. Compliers ($K = c$) followed the same treatment level that they were assigned to. Always-takers ($K = at$) took the treatment level regardless of the original assignment, and never-takers ($K = nt$) took the control level. The proportion of each subpopulation was one

of the examined factors, as described below. All subjects' values on the treatment-taken variable (*D*) was recorded as $D = 1$ for taking the treatment level and $D = 0$ the control level. The relationship of *K*, *Z*, and *D* can be expressed as

$$D = \begin{cases} 1 \mid Z = 1, K = c \text{ or } at \\ 0 \mid Z = 1, K = nt \\ 1 \mid Z = 0, K = at \\ 0 \mid Z = 0, K = c \text{ or } nt \end{cases}. \tag{66}$$

For $K = at$ or nt, the effect of the treatment assignment variable *Z* was zero.

Longitudinal data were generated following the structure of the model described in Figure 9 and Equation 63. There were four measurement points ($T = 1, 2, 3,$ and 4). The mean vectors ($\boldsymbol{\mu}_{kz}$) and covariance matrices ($\boldsymbol{\Sigma}_{kz}$) took the form presented in Table 4. For the research question involving statistical power, only non-zero γ_c conditions were involved. For the research question involving type I error rate, only zero γ_c conditions were involved.

The error variance-covariance matrix of the four measurement points had a structure where there was no covariation between any two time points. The magnitude of the error variances was determined by $R^2_{Y_{ik\_T}}$, where

$$R^2_{Y_{ik\_T}} = 1 - \frac{\sigma^2_{e\_Tk}}{\sigma^2_{Y_{ik\_T}}} = 1 - \frac{\sigma^2_{e\_Tk}}{\sigma^2_{\text{Int}\_k} + 2\lambda_{0T}\lambda_{1T}\sigma_{\text{Int\_Slp}\_k} + \lambda^2_{1T}\sigma^2_{\text{Slp}\_k} + \sigma^2_{e\_Tk}}. \tag{67}$$

$R^2_{Y_{ik\_T}}$ was the reliability of the measurement *Y* at time *T* for subject *i* from compliance class *k*. The true mean intercept (α_{Int_k}), true mean growth rate (α_{Slp_k}) and their variances ($\sigma^2_{\text{Int}\_k}$ and $\sigma^2_{\text{Slp}\_k}$) were set differently as part of the research design. Therefore, error variances also varied, but they took values with the restriction of the R^2_T vector, which was the same across compliance statuses. As different

research questions required different sets of parameter values, detailed parameter choices are discussed below.

3.1.Replications

As described below, this study involved 1,440 different configurations in total. With such a wide selection of simulation conditions, the author tried to reach a balance between the precision of the findings and keeping the study under a manageable scale. With only 100 replications for each configuration, the total computation time exceeded 72 hours with multithread computing using a powerful server. As a result, the author decided to use 1,000 replications. This replication number is consistent with similar simulation studies by Jo (2002) and Fan and Fan (2005). More rationale is provided in the Result section.

3.2.Simulation Factors

Six main factors that might affect the CACE estimation were examined: sample size [*n*], compliance composition [*PC*], effect size [*d*], reliability of measurements [*rel*], mean distances [*md*], and noncomplier-complier Level 2 covariance ratio [*var*]. Each factor contained several levels so that a variety of research scenarios were covered.

3.2.1. Sample size

The estimation of CACEs is asymptotically accurate for both the Standard IV and the MMB methods. In other words, researchers can only trust their results with a large *n*. However, in most longitudinal studies, subjects are expensive to recruit, and attrition during the study span also further reduces the number of usable study units. As a result, it is extremely difficult to have an ideal *n* for all longitudinal experiments.

For example, the JHU PIRC study had only 313 subjects. The author did not find literature providing guidance on choosing proper *n* for the CACE estimation. Therefore, the present study used a selection of *n* levels ranging from small to big in order to understand the relation between estimation success rate, bias, power, type I error rate and sample size for both estimation approaches.

Previous studies on latent growth mixture modeling (LGMM) suggested that in order to achieve reliable results in estimation, a small sample size (n = 50) only worked with big separation of latent means (i.e., standardized difference = 4 or 5) (Tolvanen, 2007). In addition, when the model was specified correctly, small sample size was found to lead to large standard errors. The situation deteriorated with the combination of lower measurement reliability. Parameter standard errors decreased to 31% of the original values when sample size increased from 50 to 100 and 44% when the sample size was 500 (Tolvanen, 2007). As the present study also examined a few other factors that could influence the estimation results together with sample size, small *n* levels were included to investigate the estimation quality while other factors were more favorable. As a result n = 50, 100, 200, 500 or 1,000 were used.

3.2.2. Compliance composition

Compliance composition is a critical factor for the CACE estimation. Earlier studies warned researchers that a weak instrument (i.e., low complier rate) could severely distort the CACE estimation (Angrist et al., 1996; Imbens & Rubin, 1997b). Using the lowest complier proportion of 30%, Jo's (2002) simulation study showed that with the presence of noncompliance, researcher should plan an even bigger sample size to compensated the influence of noncompliance. In other applied studies, the compliance rate of 30%, however, was not low enough. Imbens and Rubin's

(1997b) study used a sample size of 1,000 to compensate for a low compliance rate of 10%. Their real data analysis with Angrist and Krueger's (1991) data has an incredibly large sample size (n = 162,515), but the compliance rate was only around 2%. No simulation study has investigated the performance of the methods in the low compliance rate scenario, so it was unclear how the CACE estimation would behave. Therefore, the current study examined four levels of compliance rate, PC = 10%, 30%, 50%, and 80%, to cover low to high compliance rates. Always-takers and never-takers were set to have the same proportions.

As designs for effect size, measurement reliability, mean distance, and noncomplier-complier Level 2 covariance ratio are all inter-related, the following sections uses symbols for the true values of different parameters. A summary table is provided in the end.

3.2.3. Effect size

As mentioned earlier, the treatment assignment variable Z has no effect on always-takers and the never-takers. Therefore, setting the effect size of Z on the compliers is, in fact, setting the CACE. Jo (2002) demonstrated in her study that d was crucial for the CACE estimation: with a large effect size (0.8), empirical power could reach 0.8 with any sample size from 200 to 500 if PC = 0.5. However, it was not clear how d would influence the estimation of longitudinal CACE.

Therefore, similar to Jo's (2002) study, Cohen's d was used to measure the treatment effect size. In the form of an equation, $d = \frac{\gamma_c}{\sqrt{\sigma^2_{Slp\_c}}} = \frac{\alpha_{Slp\_c1} - \alpha_{Slp\_c0}}{\sqrt{\sigma^2_{Slp\_c}}}$, and $\sigma^2_{Slp\_c}$ was the variance of compliers' latent slopes. Values of d covered 0, 0.2, 0.5, and 0.8, representing no effect, weak effect, medium effect, and strong effect (Cohen,

1988; Cohen, 1992). When $d \neq 0$, γ_c was determined by the product of *d* and $\sigma^2_{Slp\_c}$. $\sigma^2_{Slp\_c}$ is discussed more in 3.2.6. The probability of mistakenly rejecting the null hypothesis when $d = 0$ is the type I error rate, and the probability of correctly rejecting the null hypothesis when $d \neq 0$ is the statistical power. Therefore, for research question 3, the discussion should be based on designs with $d \neq 0$; while for research question 4, $d = 0$.

3.2.4. Reliability of measurements

The earlier section mentioned that reliability of the observed variable *Y* at time *T* for subject *i* from compliance class *k* ($R^2_{Y_{ik\_T}}$) was equal to $1 - \frac{\sigma^2_{e\_Tk}}{\sigma^2_{\text{Int}\_k} + 2\lambda_{0T}\lambda_{1T}\sigma_{\text{Int\_Slp}\_k} + \lambda^2_{1T}\sigma^2_{\text{Slp}\_k} + \sigma^2_{e\_Tk}}$. $R^2_{Y_{ik\_T}}$ took the same value for all four measurement points and for all compliance classes (*rel* was used hereafter). The rationale for constraining the reliability to be the same across all measurement points was to control the ratio of the variance of random errors ($\sigma^2_{e\_Tk}$) to the variance from the latent intercept and growth ($\sigma^2_{\text{Int}\_k} + 2\lambda_{0T}\lambda_{1T}\sigma_{\text{Int\_Slp}\_k} + \lambda^2_{1T}\sigma^2_{\text{Slp}\_k}$) to be the same. Following the previous practice (Tolvanen, 2007), this study examined situations with low reliability (*rel* = 0.5) and high reliability (*rel* = 0.8).

3.2.5. Mean distance

Setting distances among mean latent intercepts and mean latent slopes of the three compliance classes was similar to setting the effect sizes. Cohen's *d* was used to measure the mean distances. Means for always-takers were determined relative to the complier assigned to the treatment level and means for the never-takers were determined relative to the compliers assigned to the control level. The mean intercepts

were the same for compliers assigned to the treatment level and the control level, so it is denoted as α_{Int}. The pooled variance was used so that two groups' variances were both involved. The following equations display more details.

$$\begin{aligned}
\alpha_{\text{Int}_at} &= \alpha_{\text{Int}_c1} + md \times \sqrt{\sigma^2_{\text{Int}_pooled}} = \alpha_{\text{Int}} + md \times \sqrt{\frac{\sigma^2_{\text{Int}_c} + \sigma^2_{\text{Int}_at}}{2}}, \\
\alpha_{\text{Int}_nt} &= \alpha_{\text{Int}_c0} - md \times \sqrt{\sigma^2_{\text{Int}_pooled}} = \alpha_{\text{Int}} - md \times \sqrt{\frac{\sigma^2_{\text{Int}_c} + \sigma^2_{\text{Int}_nt}}{2}}, \\
\alpha_{\text{Slp}_at} &= \alpha_{\text{Slp}_c1} + md \times \sqrt{\sigma^2_{\text{Slp}_pooled}} = \alpha_{\text{Slp}_c1} + md \times \sqrt{\frac{\sigma^2_{\text{Slp}_c} + \sigma^2_{\text{Slp}_at}}{2}}, \text{ and} \\
\alpha_{\text{Slp}_nt} &= \alpha_{\text{Slp}_c0} - md \times \sqrt{\sigma^2_{\text{Slp}_pooled}} = \alpha_{\text{Slp}_c0} - md \times \sqrt{\frac{\sigma^2_{\text{Slp}_c} + \sigma^2_{\text{Slp}_nt}}{2}}
\end{aligned} \tag{68}$$

md values were set to be 0.2, 0.5, and 0.8 to represent a small, medium, and large mean distance (Cohen, 1988; Cohen, 1992). Variances of different groups are discussed more in the next section

3.2.6. Noncomplier-complier Level 2 covariance ratio

When setting differences among variances of latent intercepts and slopes of the three compliance classes, three conditions were considered. Using the Level 2 covariance matrix of the compliers as the anchor, the Level 2 covariance matrices of the always-takers and the never-takers were set to be half, the same, or twice (*var* takes 0.5, 1, or 2) of that of the compliers. Always-taker and never-takers have the same Level 2 covariance matrices.

In summary, most true values for these parameters depended on the values of *d*, *md*, *rel*, and *var* and the true parameters selected for compliers assigned to the control level (the *c0* group). The mean vector and Level 2 covariance matrix of the *c0* group were set as the same as the values obtained from the real data analysis of Jo and

Muthén's (2003) study, where the equivalent complier group, the high compliance group, had a mean vector and a Level 2 covariance matrix as

$$\begin{pmatrix} \alpha_{\text{Int\_}c0} = 4.682 \\ \alpha_{\text{Slp\_}c0} = 0.118 \end{pmatrix} \text{ and } \begin{bmatrix} \sigma^2_{\text{Int\_c0}} = 0.936 & \\ \sigma_{\text{Int\_Slp\_c0}} = -0.048 & \sigma^2_{\text{Slp\_c0}} = 0.064 \end{bmatrix}.$$

Table 5 provides a summary of how the four factors determine the true parameter values used for data generation. As a result, there were 1,440 different configurations (5[*n*]*4[*PC*]*4[*d*]*2[*rel*]*3[*md*]*3[*var*]) for data generation, and for each configuration, 1,000 replications were conducted. R studio (2009-2017) was used to generate data with these models. Specifically, the "mvrnorm" function of the "MASS" package was used to generate the multivariate normal data.

Table 5

Population Matrix Designs with Respect to the Four Factors

	Never-taker		Complier		Always-taker	
	Z=0	**Z=1**	**Z=0**	**Z=1**	**Z=0**	**Z=1**
α_{Int}	$4.682 - md \times \sqrt{\frac{0.936 + 0.936var}{2}}$		4.682	4.682	$4.682 + md \times \sqrt{\frac{0.936 + 0.936var}{2}}$	
α_{Slp}	$0.118 - md \times \sqrt{\frac{0.064 + 0.064var}{2}}$		0.118	$0.118 + d \times \sqrt{0.064}$	$0.118 + d \times \sqrt{0.064} + md \times \sqrt{\frac{0.064 + 0.064var}{2}}$	
$\begin{bmatrix} \sigma^2_{\text{Int}} & \\ \sigma_{\text{Int\_Slp}} & \sigma^2_{\text{Slp}} \end{bmatrix}$	$var \times \begin{bmatrix} 0.936 & \\ -0.048 & 0.064 \end{bmatrix}$		$\begin{bmatrix} 0.936 & \\ -0.048 & 0.064 \end{bmatrix}$		$var \times \begin{bmatrix} 0.936 & \\ -0.048 & 0.064 \end{bmatrix}$	
$\begin{bmatrix} \sigma^2_{e\_T1} \\ \sigma^2_{e\_T2} \\ \sigma^2_{e\_T3} \\ \sigma^2_{e\_T4} \end{bmatrix}$	$\begin{bmatrix} \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 0 \times (-0.048) + 0^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 1 \times (-0.048) + 1^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 2 \times (-0.048) + 2^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 3 \times (-0.048) + 3^2 \times 0.064) \end{bmatrix}$		$\begin{bmatrix} \frac{1-R^2}{R^2} (0.936 + 2 \times 1 \times 0 \times (-0.048) + 0^2 \times 0.064) \\ \frac{1-R^2}{R^2} (0.936 + 2 \times 1 \times 1 \times (-0.048) + 1^2 \times 0.064) \\ \frac{1-R^2}{R^2} (0.936 + 2 \times 1 \times 2 \times (-0.048) + 2^2 \times 0.064) \\ \frac{1-R^2}{R^2} (0.936 + 2 \times 1 \times 3 \times (-0.048) + 3^2 \times 0.064) \end{bmatrix}$		$\begin{bmatrix} \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 0 \times (-0.048) + 0^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 1 \times (-0.048) + 1^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 2 \times (-0.048) + 2^2 \times 0.064) \\ \frac{1-R^2}{R^2} \times var \times (0.936 + 2 \times 1 \times 3 \times (-0.048) + 3^2 \times 0.064) \end{bmatrix}$	

3.3.Estimation

For each simulated dataset, both the Standard IV (Figure 10) and the MMB (Figure 9) methods were used to estimate the longitudinal CACE (γ_c) using Mplus 8.1 (Muthén & Muthén, 1998-2018).

As examining the effect of *var* was one goal of the present study, there were one third of the simulation designs (when *var* = 1) having homogeneous latent covariance matrices ($\sigma^2_{\varsigma k}$) and therefore homogeneous error covariance matrices ($\sigma^2_{\varepsilon\_k}$) across the three compliance classes, and the other two thirds (when *var* = 0.5 or 2) were designed to be heterogeneous on these matrices.

Fundamentally, the Standard IV approach did not use the latent class framework, so there was no need to consider covariance matrix heterogeneity when applying the Standard IV model. However, with the MMB estimation approach, one can specify the estimation models either by constraining $\sigma^2_{\varsigma k}$ and $\sigma^2_{\varepsilon\_k}$ to be the same across different compliance classes or by allowing the two matrices to be freely estimated. One would assume that the correctly specified model would perform better than the incorrectly specified model. However, previous studies have shown that a model with more parameters to estimate was also more prone to estimation failure, and separately estimating the covariance matrices for each latent class would yield unreliable estimations of higher order moments (Imbens & Rubin, 1997b; Tolvanen, 2007). Therefore, it was valuable to evaluate the estimation result with both the more parsimonious model by incorrectly constraining the covariance matrices across the

three latent classes and the more complicated model where the covariance matrices were freely estimated.

As a result, this study used three estimation methods for each simulated dataset: two MMB estimation approaches and one Standard IV approach.

1) The Standard IV approach [SIV]: All parameters were estimated without separating classes.
2) The MMB No Constraint approach [MMB-NC]: All parameters, α_{Int}, α_{Slp}, γ, σ_{ζ}^2, and σ_{ε}^2 were freely estimated within the three compliance classes. It was the correct model for all datasets but unnecessary when compliers and non-compliers had the same Level-2 covariance matrices ($var = 1$).
3) The MMB Full Constraint approach [MMB-FC]: α_{Int}, α_{Slp}, and γ were freely estimated in the three compliance classes, but σ_{ζ}^2 and σ_{ε}^2 were constrained to be equal across the three classes. It was an incorrect model when compliers and non-compliers had different Level-2 covariance matrices ($var = 0.5$ or 2) and correct model when compliers and non-compliers had the same Level-2 covariance matrices.

For each estimation using the two MMB approaches, Mplus commands "STARTS 500 20" and "STITERATIONS = 20" were used. The first command requested Mplus to generate 500 sets of different random starting values for all parameters and to do 20 iterations of the maximization on all 500 sets. The second command then made use of the 20 sets of parameter estimates that had the best

likelihood values obtained from the first stage as the new starting values. It was possible to reduce the number of estimation problems by increasing these numbers, but the computation time would increase too. The influence of changing these parameters on estimation was out of the scope of the current study, so only "STARTS 500 20" and "STITERATIONS = 20" were employed.

3.4.Evaluation Criteria

Six dependent variables were used to address the five research questions. The first dependent variable was used to answer Research Question 1, the second to the fifth dependent variables were for Research Question 2, and the sixth dependent variable was used for Research Question 3 and 4. The comparison of the two estimation methods, Research Question 6, was conducted using all six variables.

The first dependent variable, the "Success Indicator", conveyed the information about whether an estimation was successful or not. A successful estimation was defined as a converged estimation without any untrustworthy information.

Estimations can fail or yield untrustworthy results due to various reasons, such as local maxima, non-positive definite variance estimates, and sample sizes not big enough for proper parameter estimations. In such cases, Mplus would generate a warning or an error message. If failed or untrustworthy results happen in a real data analysis, researchers have to make some changes—increasing the sample size, changing the starting values, increasing the iteration times, etc.—to fix the estimation errors. Sometimes it is unfixable, especially when the sample size is too small. In the current study, the post-estimation model revision was not part of the current research interest; therefore, only estimations free of warning or error messages were defined as

"successful estimations". Consequently, the "Success Indicator" was defined as a dichotomous variable with three possible values: 1 = Successful Estimation, 0 = Non-Successful Estimation, and NA = Excluded Datasets (see 3.5.1). All cases with "NA" were excluded from all analyses. This rule also applied to all other dependent variables.

Research Question 3 and 4 were essentially about parameter significance, and parameter significance could be affected by both parameter estimation and standard error estimation. In the current study, all effect sizes were set to be equal to or greater than zero. Therefore, parameter overestimation and standard error underestimation would inflate rejection rate, and parameter underestimation and standard error overestimation would deflate rejection rate. As a result, in order to fully investigate statistical power and type I error rate, both parameter estimation bias and standard error estimation bias were included. Therefore, the second and third variables addressed the bias in parameter estimation, and the fourth and fifth variables were used to quantify the bias in standard error estimation.

The second variable was defined as the "Simple Estimation Bias", which was simply the difference between the estimated parameter and the true parameter, $\hat{\gamma}_{c\_i} - \gamma_c$, where γ_c was the true treatment effect of the *i*-th sample and $\hat{\gamma}_{c\_i}$ was the estimation of γ_c using the *i*-th sample. The third variable was the "Relative Estimation Bias", $\frac{\hat{\gamma}_{c\_i} - \gamma_c}{\gamma_c} \times 100\%$. It quantified the deviation of the estimated parameter from the true parameter, relative to the true parameter. The multiplication of the "100%" converted the bias to a percentage scale.

For both measures, a positive value indicated that the *i*-th estimation was an overestimation, and a negative value indicated an underestimation. The absolute values of the two measures were the magnitudes of the estimation bias. Both variables were continuous, and NA represented removed datasets or non-successful estimations.

Both the simple and relative bias measures are widely used in simulation studies, but both measures have their disadvantages. The simple bias is unscaled so the results are not comparable across different studies. Conclusions from one study are hardly meaningful for another, if only simple bias is used. The relative bias, however, requires the denominator (i.e., the true γ_c) to be non-zero. Because $d = 0$ was one design of the current study, $\gamma_c = 0$ was an inevitable condition. As a result, both simple and relative measures were used for the estimation bias analysis, and all designs with $d = 0$ were not included for the Relative Estimation Bias analyses.

Similar to the parameter estimation bias, there were two standard error bias measures, the "Simple Standard Error (SE) Bias" and the "Relative SE Bias". The Simple SE Bias was defined as $\widehat{SE}(\hat{\gamma}_c)_i - SE(\hat{\gamma}_c)$, where $\widehat{SE}(\hat{\gamma}_c)_i$ was the estimated standard error of parameter $\hat{\gamma}_c$ using the *i*-th replication sample and $SE(\hat{\gamma}_c)$ was the empirical standard error of parameter $\hat{\gamma}_c$. The empirical standard error was used instead of the true population standard error because the true population standard error could not be set the same way as the true population treatment effect. However, within each design cell, there were up to 1,000 replication samples (if no removed datasets or non-successful estimations) generated with exactly the same population configurations. The large number of samples formed an empirical sampling

distribution of the estimated parameter, and the standard deviation of this distribution resembled the population standard error, which was hence called the "empirical standard error".

Naturally, the "Relative SE Bias" was defined as $\frac{\widehat{SE}(\hat{\gamma}_c)_i - SE(\hat{\gamma}_c)}{SE(\hat{\gamma}_c)} \times 100\%$. Similarly, the Simple SE Bias and the Relative SE Bias could be positive, indicating overestimation, and negative, indicating underestimation. They were continuous with NA representing removed datasets or non-successful estimations.

The last dependent variable was the "Significant Indicator", suggesting whether an estimated treatment effect was statistically significant at 0.05 level. This variable has three possible values: 1 = Significant, 0 = Non-Significant, and NA = removed datasets or non-successful estimations. This variable was used for both Research Question 3 and 4. When discussing statistical power, only simulation designs with $d \neq 0$ should be included, and when examining type I error rate, only designs with $d = 0$ should be included.

3.5.Result Analysis

3.5.1. Excluded datasets

This study generated 1,440,000 datasets using 1,440 design configurations [$5_{\text{SampleSize}}$*$4_{\text{ComplierProportion}}$*$4_{\text{EffectSize}}$*$2_{\text{MsurementReliability}}$*$3_{\text{MeanDistance}}$*$3_{\text{Noncomplier-ComplierLevel2CovarianceRatio}}$]. Each configuration was a unique combination of the different levels of the six factors. These unique combinations or configurations are sometimes referred to as "design cells" in this paper.

Among the 1,440,000 generated datasets, there were 6,075 datasets ending up with no always-takers in the $Z = 0$ group (i.e., $D = 1$ and $Z = 0$), 6,043 datasets having no never-takers in the $Z = 1$ group (i.e., $D = 0$ and $Z = 1$), and 367 of the 11,751 datasets including neither. The MMB method requires to specify the number of compliance groups correctly. These 11,751 (0.8%) datasets nonetheless suggested an incorrect number of compliance groups, so they were removed from all analyses. For example, if a dataset had no always-takers in the $Z = 0$ group, one would assume that only never-takers and compliers were in the population and would incorrectly use an MMB model with only two compliance groups. However, the truth was that there was in fact three groups (because of the simulation design), and no always-takers in the $Z = 0$ group only happened because of random chance of sampling. In this case, the issue became incorrect specification of model and was not the research interest of the current study.

3.5.2. Analysis with factorial ANOVA

Simulation studies normally include a few categorical factors each having several levels, so the available information rapidly becomes overwhelming. Many researchers recommended using inferential statistics in analyzing simulation results (Bandalos & Gagné, 2012; Harwell, Rubinstein, Hayes, & Olds, 1992; Hauck & Anderson, 1984, Skrondal, 2000) to tease out peripheral information and therefore better focus on the core findings. Specifically, using a factorial ANOVA analysis to pinpoint the contribution of each factor, quantified by η^2 or partial η^2, was widely accepted by various researchers (Fan & Sivo, 2007; Bandalos & Gagné, 2012).

Therefore, in the current study, the factorial ANOVA procedure was employed as the preliminary analysis tool in selecting important factors. With the guidance of the ANOVA results, more detailed descriptive statistics and graphical techniques were followed for further investigation.

Independent variables. All simulation factors, *n*, *PC*, *d*, *rel*, *md*, and *var*, were used as the independent variables for the factorial ANOVA analyses. They were all categorical. For certain analyses, some factors or factor levels were dropped. For example, when investigating Relative Estimation Bias, $d = 0$ was dropped because only datasets with none-zero Effect Size were included for analyses.

Modeling. As some of the dependent variables were dichotomous and some were continuous, a generalized factorial ANOVA approach was applied. For continuous dependent variables, a linear regression model was used first, and for dichotomous dependent variables, a logistic regression model first. These models all included the main effect terms of the six independent variables and all of their 2- and 3-way interaction terms. More than 3-way interactions were not included because high-way interactions are usually too complicated to interpret.

For linear regression models, an analysis of variance was followed to calculate the variance explained by each term, and a partial η^2 (in percentage form) was calculated for each term, $\text{Partial } \eta^2 = \frac{SS_{\text{A}}}{SS_{\text{A}} + SS_{\text{Error}}} \times 100\%$, where SS_A was the sum of squares for factor A and SS_{Error} was the error sum of squares (Cohen, 1973).

The reason for using the partial η^2 instead of the non-partial one was that subsequently each term's partial η^2 was used to determine its importance by applying Cohen's (Cohen, 1988) rule of thumb for one-way ANOVA. Cohen (1988) defined that in one-way ANOVA, $\eta^2 = 2\%$ suggested a "small" effect, 6% "medium", and 14% "large". While in factorial ANOVA with more than one predictor, the partial η^2 is a closer approximation of the one-way ANOVA η^2, and the rule of thumb can thus readily be applied.

The incremental type I sum of squares were used to calculate all partial η^2s. When using the type I sum of squares, each effect is added into the model in a sequential order where later added terms are conditional on earlier added terms. For example, if a model contains factor A, factor B, and their interaction term AB, and the three terms are entered into the model by the order of A, B, and AB, using type I sum of squares, SS_A is the sum of squares for factor A, SS_B is the added sum of squares for factor B conditional on the model already including factor A, and SS_{AB} is the added sum of squares for the interaction term AB conditional on the model already including factor A and factor B.

With a balanced factorial design (the original design of this study), the factor entering order would not change the estimation of the sum of squares. However, due to the deletion of datasets because of the aforementioned reason (section 3.5.1), no always-takers in the $Z = 0$ group and/or no never-takers in the $Z = 1$ group, and the fact that non-successful estimations (see more in section 4.2) were also excluded from the final analyses, the final results did not preserve the balanced design. To

compensate for this downside, all regression models were run six times in order to rotate the entering order of the six factors. For example, the first model would enter the six factors in the order of *n*, *PC*, *d*, *rel*, *md*, and *var* (and their 2-way and 3-way interaction terms). The second model then would enter the six factors in the order of *PC*, *d*, *rel*, *md*, *var*, and *n* (and their 2-way and 3-way interaction terms) by moving the last term in the earlier model to be the first term in the current model. The same rule was applied to all ANOVA analyses.

For each model, the partial η^2 was calculated for all terms. In the end, six partial η^2s were obtained for each term. There might be great variation among the six partial η^2s, but if the maximum partial η^2 of a term was bigger than the "medium" effect size, 6%, that term was used for further investigation with tables and figures.

For logistic regression models, an analysis of deviance was followed to calculate the deviance reduced by each term. A pseudo-partial η^2 adapted from McFadden's (1973) R^2 to approximate the calculation of the partial η^2 in linear regression was calculated for each term,

$$\text{Pseudo Partial } \eta^2 = \frac{D_{\text{A}}}{D_{\text{A}} + D_{\text{Residual}}} \times 100\%$$
$$= \frac{D_{\text{A}}}{D_{\text{A}} + (D_{\text{Null\_Model}} - D_{\text{Applied\_Model}})} \times 100\%,$$

where D_{A} was the deviance reduced by adding factor A to the model and D_{Residual} was the deviance difference between the null model and the model applied. D_{A} was also calculated in the same way as the type I sum of square and therefore was subject to the entering order of the independent variables.

Note that only 10% of the result data were used for any logistic regression analysis because of the huge demand for computer memory and extremely lengthy computation time. In order to check if 10% of the total data were able to yield a reliable estimation of the pseudo partial η^2 for each term, the analysis process was applied to three random samples. Each sample contained results from 10% of the total datasets, and the three samples had no overlap. For each sample, the six rotation procedure used for the linear regression model were implemented too. Therefore, for each term used for each sample, there was a maximum pseudo partial η^2. Subsequently, the mean of the three maximum pseudo partial η^2s was also calculated as the "Mean of All Maxes" for each term. All terms with their "Mean of All Maxes" value bigger than 6% were kept.

The next step compared the three samples' kept results to uncover potential discrepancies introduced by using only 10% of the data. If there was great variation among the three samples' kept results, the process would start again with another three samples each containing a higher percentage of the total data (e.g., 20%). This process would not stop until the sample variation diminished. The rest of the procedure was the same as what was described in the linear regression section.

Following the guidance of the factorial ANOVA results, descriptive statistics tables and plots were engaged for further examination. Specifically, all main effects were further examined with a table presenting the level means of the six factors and a figure graphically displaying these level means. For any selected interaction term, a table and a visual aid figure were included for further examination.

Last but not least, one important goal of simulation studies is to provide practical guidance for researchers and practitioners, especially on controllable factors. For example, simulation studies can answer questions such as "What is the minimum sample size in order to reach a satisfactory estimation success rate by using this method?" or "How reliable my measurements should be so that the parameter bias is still acceptable?" Among the six simulation factors used in this study, only Sample Size and Measurement Reliability were manipulable for researchers. Hence, for each research question, this study also included guidance on the two factors as a function of other simulation factors.

For each research question, a cross tabulation table was created only using factors identified by the factorial ANOVA analysis. The tables highlighted simulation conditions that were favorable for each research question.

For example, for Research Question 1, if the factorial ANOVA identifies that *n* and *var* are the only two important factors for successful estimations, the cross tabulation table will include three factors, *n*, *rel*, and *var*. The cross tabulation table will present all configurations across all levels of these three factors. In this imaginative case, there will be 30 (5[*n*]*2[*rel*]*3[*var*]) numbers included for the table. Each number is a summary of the 48 (4[*PC*]*4[*d*]*3[*md*]) design cells with regard to a specific combination of *n*, *rel*, and *var*. If, for instance, with $n = 50$, $rel = 0.5$, and $var = 0.1$, the number is 25%, it means that 12 out the 48 design cells have their mean successful estimation rate (across up to 1,000 replications for each cell) meeting a pre-

specified criterion. If the number is 100%, it means that all 48 design cells meet the criterion.

The criterion was chosen differently for each research question. For Research Question 1, the mean success rate was calculated, and the criterion was set to be at or above 90.91% (Gagne & Hancock, 2006). Using the example above, a number of 25% means that 12 out of 48 design cells have their mean success estimation rate higher than or equal to 90.91%.

For Research Question 2, only the Relative Estimation Bias and the Relative SE Bias variables were used, and a cross tabulation table was created for each variable. According to Muthén, Kaplan, and Hollis (1987), relative bias of 10% could be considered as "negligible". Therefore, the criterion was set to be between the bound of ±10%. In this case, a number from the two cross tabulation tables would indicate the percentage of design cells having mean bias within the bound of ±10%.

For Research Question 3, the mean significant rate was calculated for designs with $d \neq 0$. Each number represents the percentage of design cells with cell average empirical power higher than or equal to the pre-specified satisfactory power rate, 80%.

For Research Question 4, the mean significant rate was calculated for designs with $d = 0$. Each number represents the percentage of cells with cell average type I error rate within the bound of 4.5% and 5.5%. This bound was chosen because the non-biased type I error rate should be 5% as all significance tests were conducted at

the level of 0.05. Following the aforementioned “negligible” rule of bias, a 10% interval was constructed around 5%, i.e., 4.5% to 5.5% (Bradley. 1978).

Chapter 4: Results

This chapter first provides the rationale for using 1,000 replications for each design cell. The rest of the chapter is organized by the order of the research questions. Sections 4.1 and 4.2 address the first two questions. Section 4.3 discusses Research Question 3 and 4 together because power and type I error are both related to significance rate. Research Question 5 is addressed across the three sections where the two estimation methods are compared with respect to each criterion.

4.1. Replication Check

As mentioned in Chapter 3, for each design cell, the author generated 1,000 replications. In order to examine if 1,000 replications was enough to yield reliable statistics for each cell, plots trying to depict the estimation of the six key variables along the increase of replication time were created.

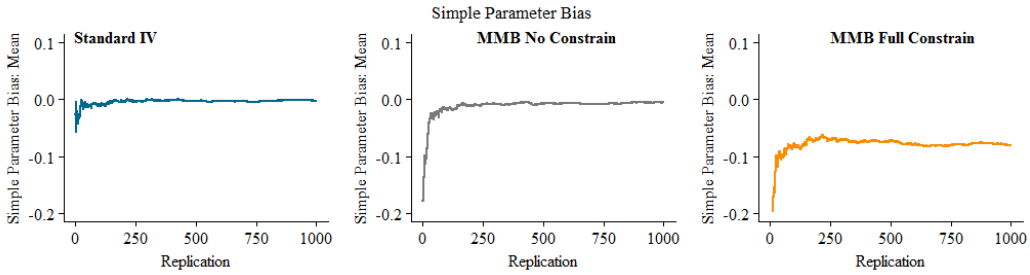

Figure 11. Mean value of Simple Parameter Bias as a function of replication time for the three estimation methods

Figure 11 is an example of these plots. This figure shows the mean value of Simple Parameter Bias as a function of replication time for the three estimation methods for the design cell with $n = 500$, $PC = 0.3$, $d = 0$, $rel = 0.5$, $md = 0.2$, and $var = 0.5$. With the increase of replication number, the mean value started to converge.

The variation of the mean values became minimum when replication approached 1,000. This was an indication that replication of 1,000 was enough to yield reliable estimation of Simple Parameter Bias for that cell. Similar plots can be observed for other outcome variables, although the mean bias converged quicker than others.

4.2.Success Rate

This section presents the results regarding research question one: how will each of the six factors affect the estimation success rate of the Standard IV and the MMB methods? Success estimation indicator was used for the investigation.

Section 4.2.1 presents the result of the factor effects. Section 4.2.2 provides guidance on choosing sample size and reliability value. Section 4.2.3 includes further analysis explanation of the other research questions as a result of the analysis of estimation success rate.

4.2.1. Results of factor effects

Following the analysis procedure described in section 3.5, three 10% random samples were used for the six-rotation factorial ANOVA analysis. For each sample, there was one maximum pseudo partial η^2 (out of the six rotations) for each term included in the logistic regression model. As a result, each term had three maximum pseudo partial η^2s (one for each sample). For terms with their "Mean of All Maxes" (i.e., the mean of the three maximum pseudo partial η^2s) bigger than 6% were checked closely to inspect the degree of sample variation. All three samples had very similar

maximum pseudo partial η^2s for the kept terms ("Mean of All Maxes" bigger than 6%). Therefore, the procedure stopped with the three 10% samples.

Table 6

Factorial ANOVA Result: Significance Indicator

		Max Pseudo Partial Eta Squares			
Estimation Approach	**Term**	**Sample 1**	**Sample 2**	**Sample 3**	***Mean of All Maxes***
Standard IV	Sample Size	**20.67**	**20.09**	**20.55**	***20.44***
	Measurement Reliability	**9.82**	**9.23**	**9.24**	***9.43***
	Complier Proportion	1.93	1.71	2.09	*1.91*
MMB No Constraint	Sample Size	**53.57**	**53.26**	**53.35**	***53.40***
	Complier Proportion	**30.00**	**29.42**	**29.65**	***29.69***
	Measurement Reliability	**13.88**	**13.95**	**13.51**	*13.78*
MMB Full Constraint	Sample Size	**26.73**	**26.36**	**26.61**	*26.57*
	Measurement Reliability	**15.57**	**15.74**	**15.57**	*15.63*
	Sample Size*Measurement Reliability	0.89	0.92	0.88	*0.90*

Note. $\eta^2 \geq 14\%$: bold and underlined ; $6\% \leq \eta^2 < 14\%$: bold; $\eta^2 < 6\%$: grayed out.

Table 6 presents the selected terms and their three maximum pseudo partial η^2s. The last column contains the means of the three maximum pseudo partial η^2s for terms presented in this table. All terms follow the order of the last column.

Terms with the last column value meeting the 6% criterion are all presented in this table, but in order to include enough information, at least three terms are included for each estimation method (same rule applies to all other similar tables later). If a term has its "Mean of All Maxes" greater than or equal to Cohen's "large" effect size, 14%, that term's partial η^2s are bold and underlined; if greater than or equal to the "medium" effect size, 6%, but smaller than 14%, bold; if smaller than 6% but kept in the table because of "to include enough information", grayed out. Throughout this

paper, all similar summary tables use the same formatting rule. The following discussion refers to the "Mean of All Maxes" column when discussing each term.

The main effects of Sample Size and Measurement Reliability were important for all three estimation methods, and they were the only selected terms for the SIV and the MMB-FC methods. The main effect of factor *n* was the dominant term for all three estimation methods. Its pseudo partial η^2 for the MMB-NC estimation approach was as high as 53.40%, meaning the main effect of *n* accounted for 53.40% of the sum of the deviance reduced by this term plus the residual deviance. Although slightly smaller than the pseudo partial η^2 of the MMB-NC approach, the values were 20.44% for the SIV method and 26.57% the MMB-FC method. For all three estimation methods, term *n* had a "large" effect.

The main effect of factor *rel* was the second most import term for the SIV and the MMB-FC approaches, accounting for 9.43% and 15.63% of the sum of the deviance reduced by this term plus the residual deviance respectively. While for the MMB-NC approach, term *rel* was the third most influential term, taking 13.78% of the sum of the deviance reduced by itself plus the residual deviance. For all three estimation methods, term *rel* at least had a "medium" effect.

For the MMB-NC method, there was an extra term having a considerable effect on its success rate, the *PC* term. It was the second most important term and had a "large" effect with pseudo partial $\eta^2 = 29.69\%$.

The factorial ANOVA results suggested that none of the three methods was considerably influenced by any interaction terms on their success rate; therefore, the following discussion only focused on the main effects.

Table 7

Success Rate Means at each Level of the Factors for the Three Approaches

		Standard IV	MMB No Constraint	MMB Full Constraint
Overall	Overall	**93.53**	**30.33**	89.06
Sample Size	50	**81.13**	**0.04**	67.53
	100	**90.46**	**3.20**	84.60
	200	**96.14**	**21.49**	93.46
	500	**99.48**	**53.68**	99.05
	1000	**99.96**	**71.99**	99.84
Complier Proportion	0.1	**90.17**	**7.59**	87.84
	0.3	**94.63**	**38.11**	88.27
	0.5	**94.66**	**45.53**	89.12
	0.8	**94.70**	**30.07**	91.09
Effect Size	0	**93.72**	**30.29**	89.07
	0.2	**93.60**	**30.34**	89.00
	0.5	**93.48**	**30.35**	89.13
	0.8	**93.33**	**30.32**	89.05
Measurement Reliability	0.5	**89.27**	**22.07**	81.19
	0.8	**97.79**	**38.58**	96.94
Mean Distance	0.2	**93.40**	**28.09**	88.70
	0.5	**93.49**	**30.00**	89.01
	0.8	**93.70**	**32.89**	89.48
Noncomplier-Complier Level 2 Covariance Ratio	0.5	**93.46**	**33.31**	87.31
	1	**93.79**	**27.24**	89.77
	2	**93.35**	**30.43**	90.11

Note. In the same row, the biggest value was bold, and the smallest was bold and underlined.

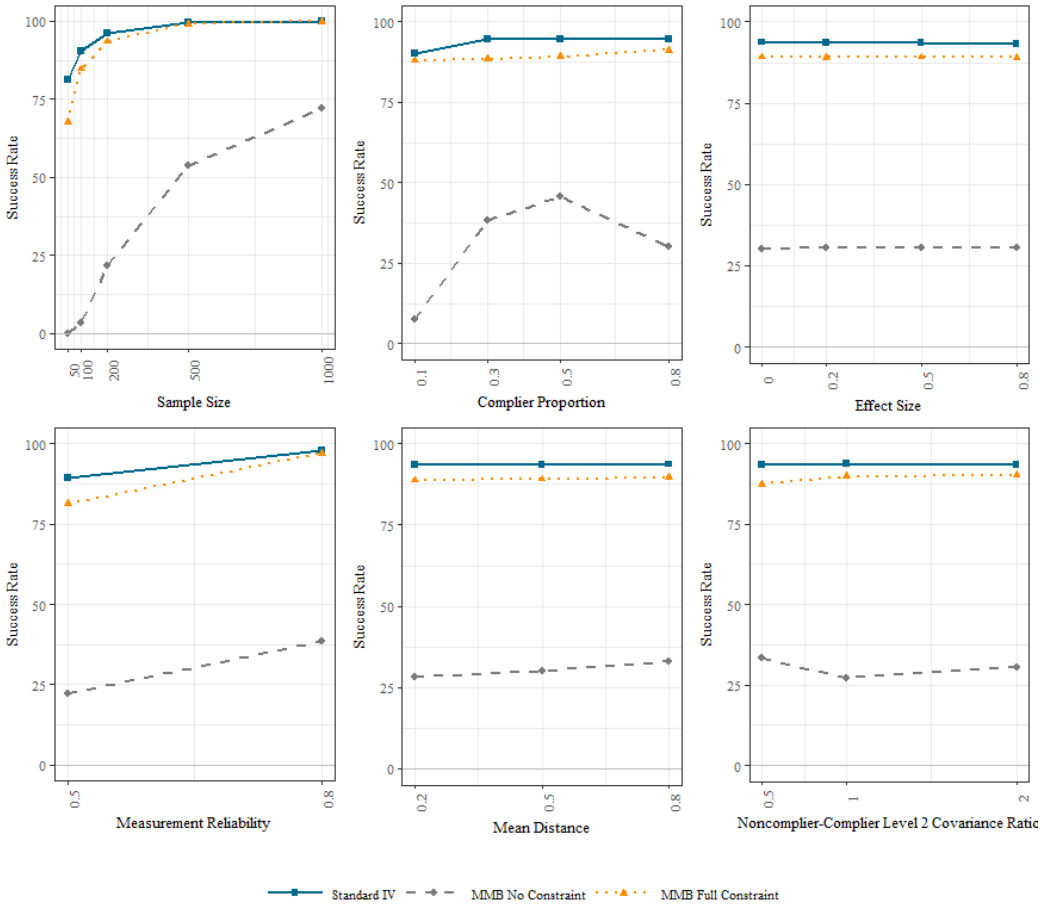

Figure 12. Success rate means at each level of the factors for the three approaches

Table 7 displays the overall success rates for the three estimation methods in the first row. The rest of the table presents the success rates with regard to each individual level of the six factors. In the same row, the highest success rate is bold, and the lowest is bold and underlined. Figure 12 shows a plot for each factor and includes the level means for all three methods in one plot. The horizontal axis of each plot demonstrates different levels of a factor, and the vertical axis depicts the mean success rate. Note that in order to preserve the interval nature of the factors, the

numbers on the horizontal axis are not equally distanced. Each plot uses three different lines to represent the three estimation methods.

Within each estimation method, it was consistent with the factorial ANOVA results. The factors selected in the factorial ANOVA analysis also showed obvious variations across their levels. A clear and positive trend was observed for the Sample Size factor for all three estimation approaches: the mean success rate went up as *n* increased. This trend was very noticeable for the MMB-NC approach—the mean success rate changed from 0.04% to 71.99% (71.93% increase) when *n* increased from 50 to 1,000. For this estimation method, the increase of the success rate was also relatively proportional to the increase of *n*. The line for this estimation method in the "Sample Size" plot approximated a straight line. While for the other two approaches, even $n = 50$ yielded a relatively high mean success rate, 81.13% for the SIV method and 67.53% for the MMB-FC method. When *n* reached 200, both methods had a success rate higher than 90%. For both estimation methods, the influence of *n* was most evident when changing Sample Size from 50 to 200. With $n > 200$, success rate approached 100%.

The upsurge in estimation success brought by changing *rel* from 0.5 to 0.8 was also sizeable: the SIV method's average success rate increased from 89.27% to 97.79% (an 8.52% increase), the MMB-FC approach increased from 81.19% to 96.94% (a 15.75% increase), and the MMB-NC approach increased from 22.07% to 38.58% (a 16.51% increase).

As mentioned in the factorial ANOVA analysis section, the factor *PC* showed an evident effect on the success rate for the MMB-NC approach only. In the "Complier Proportion" plot of Figure 12, only the line of the MMB-NC approach fluctuated remarkably across different *PC* levels while the other two lines did not change too much. The trend caused by *PC* on the success rate while using the MMB-NC approach suggested that the success rate went up when changing from low to medium *PC*. However, the success rate started to decrease when switching from medium to high *PC*. With $PC = 0.5$, the mean success rate culminated at 45.53%, while with a too high or too low *PC*, the mean success rate was lackluster (7.59% when $PC = 0.1$, 30.07% when $PC = 0.8$). For the other two estimation methods, the effect of *PC* was not too obvious, but there was a general positive trend for both methods.

In terms of the other three factors, *d*, *md*, and *var*, their influence on the success rate was too small to exhibit a noticeable pattern in Figure 12. The only patterns worth mentioning was for the MMB-NC method with *md* and *var*. Factor *md* had a very small positive effect. Factor *var* had a nonlinear effect: when $var = 1$, the mean estimation success rate was the lowest.

Comparing the three estimation methods, the SIV method (93.53%) and the MMB-FC method (89.06%) yielded a much higher overall estimation success rate than the MMB-NC method (30.33%). When looking at each individual level within the factors, the same pattern remained: the SIV method always had the highest success rate, the MMB-FC method had a slightly lower rate and the MMB-NC

method was always much lower. Only with bigger sample size or more reliable measurements, the difference between the SIV and the MMB-FC methods diminished. The improvement on only one factor, having $PC = 0.5$, also brought the success rate of the MMB-NC approach to be closer to the other two but their difference was still quite sizable. As it was shown below (Table 8), only when all three factors had the very favorable conditions, the success rate of the MMB-NC method would approach the other two.

4.2.2. Guidance on choosing *n* and *rel* with respect to estimation success rate.

Table 8

Percentages of Cells with Cell Average Success Rate Higher than or Equal to the Satisfactory Success Rate[a]

		Standard IV				MMB No Constraint				MMB Full Constraint			
n	**rel**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**
50	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	0.8	2.78	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	0.8	50.00	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	***100.00***	***100.00***	***100.00***	***100.00***
200	0.5	41.67	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	0.00	8.33	8.33	50.00
200	0.8	97.22	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	***100.00***	***100.00***	***100.00***	***100.00***
500	0.5	***100.00***	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	***100.00***	***100.00***	***100.00***	***100.00***
500	0.8	***100.00***	***100.00***	***100.00***	***100.00***	0.00	69.44	***100.00***	0.00	***100.00***	***100.00***	***100.00***	***100.00***
1000	0.5	***100.00***	***100.00***	***100.00***	***100.00***	0.00	0.00	0.00	0.00	***100.00***	***100.00***	***100.00***	***100.00***
1000	0.8	***100.00***	***100.00***	***100.00***	***100.00***	11.11	***100.00***	***100.00***	***100.00***	***100.00***	***100.00***	***100.00***	***100.00***

Note. Percentages equal to 100% were italicized, bold, and underlined.

[a]"Satisfactory Success Rate" meant that the cell average success rate was 90.91% or higher.

To provide guidance on choosing adequate sample size and measurement reliability to yield a satisfactory estimation success rate, Table 8 summarizes the

success rates with respect to the three factors identified above (i.e., *n*, *PC*, and *rel*). Within each estimation method, each number in this table is the percentage of cells with cell average success rate higher than or equal to 90.91%. For example, among the 36 cells (4[*d*]*3[*md*]*3[*var*]) with $n = 50$, $rel = 0.8$, and $PC = 0.1$, 2.78% of them (i.e., 1 cell) had cell mean success rate higher than or equal to 90.91%. Numbers equal to 100% are italicized, bold, and underlined.

For the SIV approach, when Complier Proportion was 0.1, the minimum requirements for sample size and measurement reliability were 500 and 0.5, meaning that if a sample had only 10% compliers the minimum sample size for guaranteeing a satisfactory success rate across all considered levels on *d*, *md*, and *var* was 500 and the minimum measurement reliability could be as low as 0.5. When $PC \geq 0.3$, conditions with $n = 50$ and $rel = 0.8$ or with $n = 200$ and $rel = 0.5$ would lead to a satisfactory convergence rate.

For the MMB-NC approach, when $PC = 0.1$, none of the cells reached a satisfactory convergence rate, meaning that with such a low complier proportion even sample size reaching 1,000 together with using reliable measurements would not guarantee a satisfactory convergence rate across all considered levels on *d*, *md*, and *var*. When $PC = 0.3$ or 0.8, the minimum measurement reliability required was 0.8 with a combination of a sample size of 1,000. When $PC = 0.5$, the minimum reliability was still 0.8 but the minimum sample size decreased to 500.

For the MMB-FC method, the minimum requirement for *n* was 500 when *rel* = 0.5, but a smaller *n* of 100 or 200 would also guarantee that all cells with the same configuration on *n* and *rel* meeting the satisfactory criterion if *rel* = 0.8.

4.2.3. Selecting applicable conditions for the MMB-NC approach

As discussed above, the MMB-NC estimation method had extremely restricted conditions where it could reach the satisfactory success rate. Even for the other two estimation methods, not all conditions met the satisfactory success rate. It would be rather limiting to investigate the approaches if only conditions meeting the satisfactory criterion were included. In addition, for certain factor levels, the MMB-NC approach could yield tremendously low success rate; therefore, the analysis results using these conditions could be particularly unreliable. For example, when n = 50 and PC = 0.1, only 4 out of 72,000 datasets were successfully estimated using the MMB-NC method. In this case, any statistics computed with only these four data points would be very unreliable. Therefore, in order to explore the MMB-NC approach under fairly wide settings and to obtain trustworthy analysis results, a middle ground was chosen: only conditions within which the nested cells had 300 or more usable datasets should be used for the MMB-NC approach. All 1,440 configurations included in the simulation design had 300 or more usable datasets for the other two estimation methods, so there was no need to select the "Applicable Conditions" for these two methods.

As a result, the configurations with $n \geq 500$ and $PC \geq 0.3$ were defined as the "Applicable Conditions" for the MMB-NC approach. Only under the "Applicable

Conditions" were the MMB-NC approach's results analyzed and were compared to the other two estimation approaches. Therefore, the analyses of other dependent variables included three stages. The first stage was to use all datasets to analyze the SIV and the MMB-FC methods. In this part, the effect of each factor on the two estimation approaches was identified, and the two estimation methods were compared first. The second stage was to only involve datasets under "Applicable Conditions" for the analysis of the MMB-NC method. In this part, the effect of each factor on the MMB-NC approach was identified, and the three estimation methods were compared under the "Applicable Conditions". The last stage made recommendations for choosing sample size and measurement reliability as a function of the ANOVA selected factors.

In the following discussion, only successfully estimated results are used for all analyses.

4.3.Estimation Bias

This section presents the results regarding research question two: how will each of the six factors affect the estimation accuracy (biased or unbiased estimation) of the Standard IV and the MMB methods? Simple Estimation Bias and Relative Estimation Bias were used for investigating parameter estimation bias; Simple SE Bias and Relative SE Bias were used for standard error estimation bias.

Section 4.3.1 and 4.3.2 present the result for the two types of bias respectively. Within each of these two sections, the analysis results for the SIV and the MMB-FC approaches using all conditions are discussed first, the analysis results for the MMB-

NC approach using only "Applicable Conditions" follows, and the recommendations on choosing sample size and measurement reliability are presented in the end.

4.3.1. Parameter estimation bias

Results of factor effects with all conditions. This section examined the two estimation bias measures for the SIV and the MMB-FC approaches using all simulation conditions.

Table 9

Factorial ANOVA Result: Estimation Bias Measures with the SIV and the MMB-FC Methods

		Max Partial Eta Square	
Estimation approach	**Term**	**Simple**	**Relative**
Standard IV	Sample Size*Complier Proportion	0.15	NA
	Mean Distance	0.05	0.04
	Mean Distance*Complier Proportion	0.05	NA
	Complier Proportion	NA	0.11
	Effect Size*Sample Size*Complier Proportion	NA	0.08
MMB Full Constraint	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	3.53	2.65
	Complier Proportion	1.53	1.11
	Sample Size*Complier Proportion	1.09	NA
	Effect Size*Noncomplier-Complier Level 2 Covariance Ratio	NA	1.37

Note. $\eta^2 < 6\%$: grayed out. *NA* suggested that the term was not selected or was not one of the top three terms for the dependent variable.

Table 9 presents a summary of the linear regression results using the Simple Estimation Bias and the Relative Estimation Bias as the dependent variables. This table includes the terms selected by using each dependent variable and their maximum partial η^2s obtained from the six rotations. As mentioned before, at least three partial

η^2s are presented in tables like this for each dependent variable within each estimation approach, but the ones not meeting the 6% cutoff criterion are grayed out. In this table, none of the partial η^2s is bigger than or equal to 6%, so all partial η^2s are grayed out. "NA" in this table indicates that this term is not chosen or is not one of the top three terms using the current variable but is chosen or is one of the top three term using another dependent variable. For example, in the first row, "NA" shows up in the Relative Estimation Bias column. It suggests that the term "Sample size*Complier proportion" is one of the top three terms with the biggest partial η^2s for the dependent variable Simple Estimation Bias but not one of the top three for the dependent variable Relative Estimation Bias. Within each estimation method, the terms are ordered by the "Simple" column.

After applying the 6% cutoff criterion to the maximum partial η^2s, no term was selected for either estimation approach. Especially for the SIV approach, the biggest partial η^2 was for the interaction term of "Sample Size*Complier Proportion", although it was only 0.15%, meaning that the interaction between *n* and *PC* only explained 0.15% of the sum of variance of that term plus error variance. The MMB-FC approach had a slightly better term: the interaction term of "Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio" had a partial η^2 of 3.53% when using the simple bias and 2.65% when using the relative bias. Considering that most partial η^2s were extremely small, although the partial η^2 for this interaction term was smaller than 6%, it was larger than the small effect criterion, 2%.

Therefore, an investigation of this interaction was included after the main effect examination.

Table 10 summarizes the simple and relative bias means over different levels of the six factors. Figure 13 and Figure 14 plot these means for the two dependent variables respectively. As all mean values of the simple bias are relatively small, they are rounded to four decimal points in Table 10. Note that for all relative bias analyses, datasets with $d = 0$ were removed. All relative bias means are on a percentage scale. Relative bias columns are also shaded in light gray for distinction. Within one dependent variable and one single row, the bigger absolute value was bold and the smaller absolute value was bold and underlined.

Table 10

Simple and Relative Estimation Bias Means at each Level of the Factors for the SIV and the MMB-FC Approaches

		Standard IV		MMB Full Constraint	
		Simple Bias	**Relative Bias**	**Simple Bias**	**Relative Bias**
Overall	Overall	-0.0252	-27.14	0.0213	22.86
Sample Size	50	-0.0312	-30.23	0.0684	74.89
	100	-0.0286	-31.15	0.0416	45.05
	200	-0.0279	-30.52	0.0244	25.29
	500	-0.0257	-29.18	0.0028	2.92
	1000	-0.0146	-15.86	-0.0111	-12.23
Complier Proportion	0.1	-0.0607	-64.25	0.0960	103.53
	0.3	-0.0319	-34.99	-0.0039	-4.30
	0.5	-0.0075	-8.31	-0.0045	-5.08
	0.8	-0.0019	-2.08	-0.0019	-2.09
Effect Size	0	-0.0210	NA	0.0217	NA
	0.2	-0.0228	-45.14	0.0209	41.35
	0.5	-0.0270	-21.35	0.0210	16.59
	0.8	-0.0302	-14.90	0.0216	10.65
Measurement Reliability	0.5	-0.0312	-33.79	0.0228	24.27
	0.8	-0.0198	-21.07	0.0200	21.67
Mean Distance	0.2	-0.0088	-9.45	0.0112	10.58
	0.5	-0.0259	-28.24	0.0231	25.73
	0.8	-0.0410	-43.69	0.0295	32.18
Noncomplier-Complier Level 2 Covariance Ratio	0.5	-0.0182	-20.23	-0.0830	-91.29
	1	-0.0232	-25.30	0.0302	33.38
	2	-0.0344	-35.93	0.1134	123.02

Note. In the same row and within the same bias measure, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

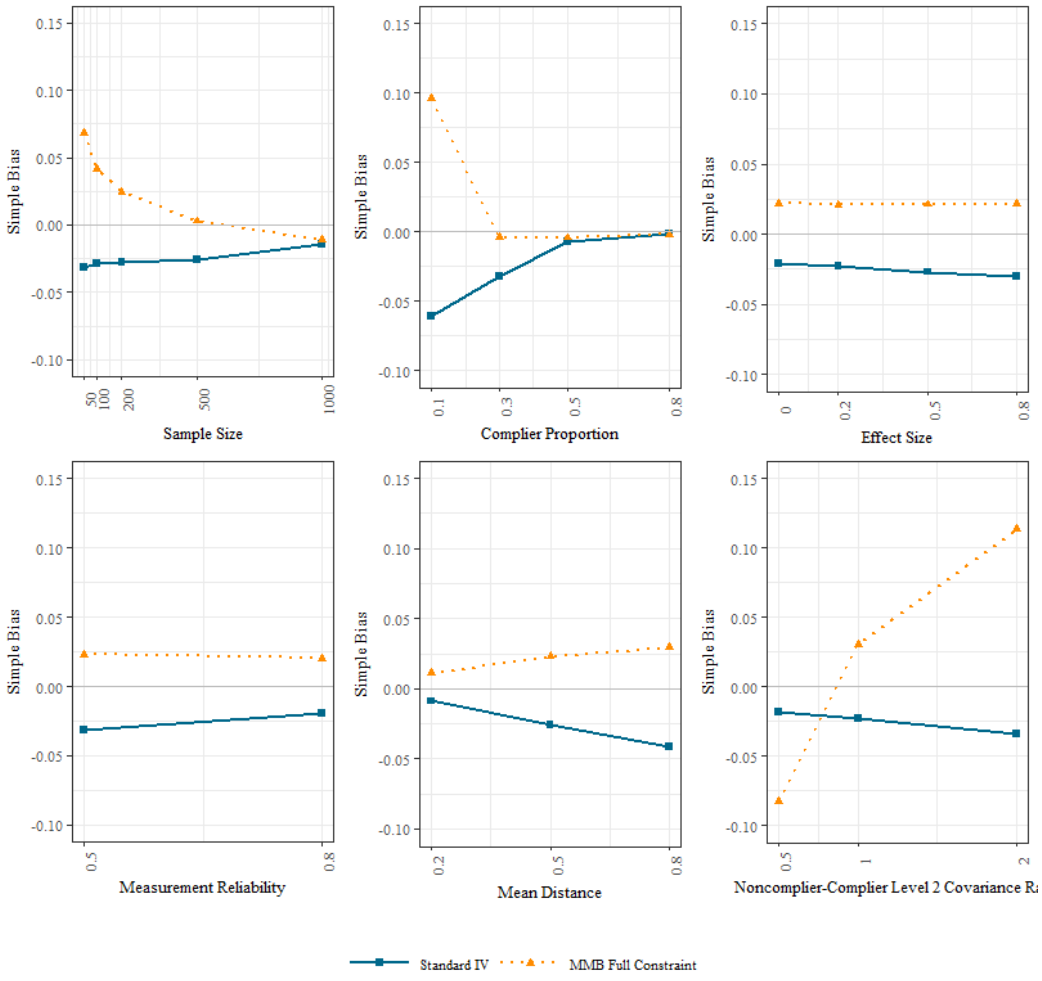

Figure 13. Simple Estimation Bias means at each level of the factors for the SIV and the MMB_FC approaches.

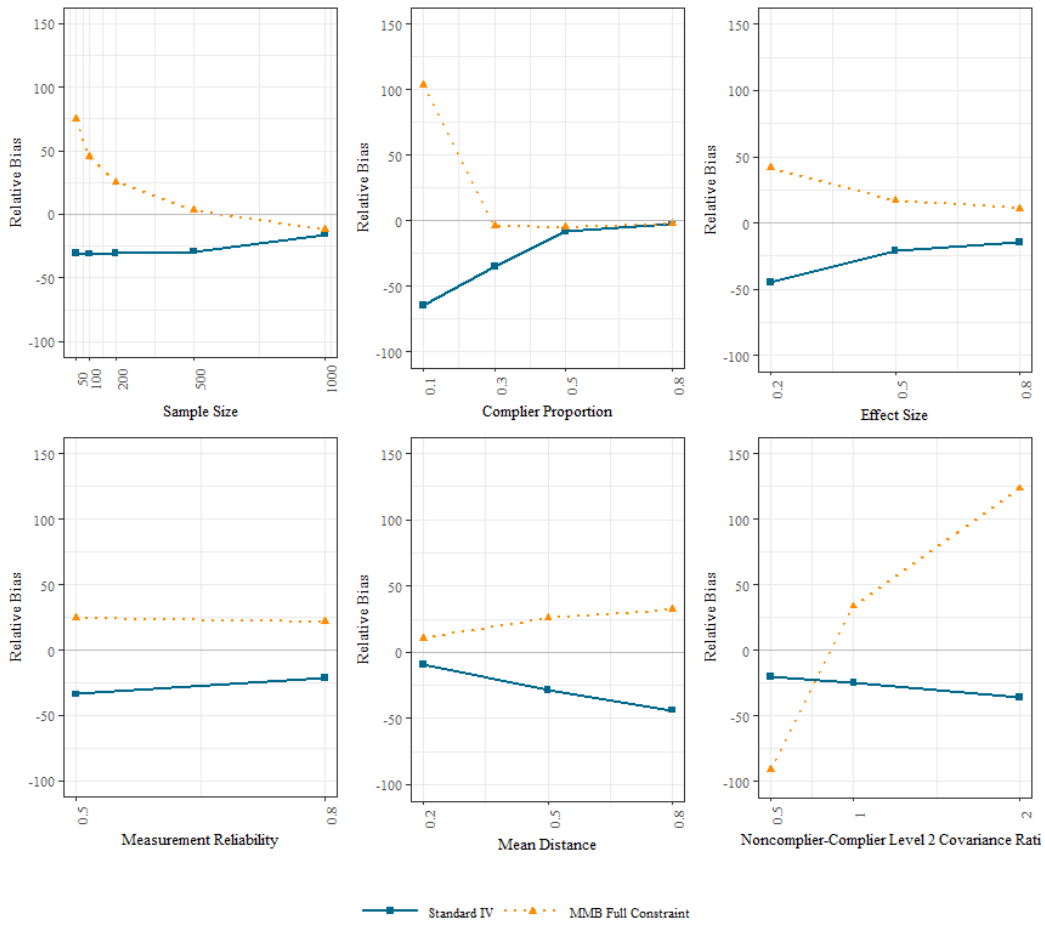

Figure 14. Relative Estimation Bias means at each level of the factors for the SIV and the MMB_FC approaches.

The trends of the two bias measures over different levels of five factors were similar to each other. The only difference was found for the Effect Size factor. Comparing to the simple bias, where *d* did not show obvious impact, the relative bias was overtly under the influence of *d*. This was true for both estimation methods. However, note that as the Effect Size factor did not show substantial impact on the simple bias, the influence of Effect Size on the relative bias variable was mainly because of using the true treatment effect as the denominator. In other words, the

estimation quality did not improve with bigger effect size because the simple bias did not change with a bigger effect size.

As the trends of the two bias measures are similar for both estimation approaches regarding the other five factors, the discussion below combines the two together. Overall, the SIV approach underestimated the treatment effect. The mean values across all conditions were −0.0252 measured by the simple bias and −27.14% by the relative bias. Among all six factors, Complier Proportion had the greatest effect on the estimation bias. The mean values started with very low negative numbers (Simple = −0.0607/Relative = −64.25%) when $PC = 0.1$. As PC increased, the mean values also increased but remained negative. At the same time, their absolute values decreased. When $PC = 0.8$, the mean values shrank to −0.0019 and −2.08% (0.0588/62.17% decrease in the magnitudes). The change in the mean values was almost proportional to the change of PC when $PC < 0.5$. There was not much change when increasing PC from 0.5 to 0.8.

A similar pattern was found with respect to another two factors, Sample Size and Measurement Reliability, but their effects were almost negligible, especially for *rel*.

Mean Distance had the second greatest effect on the estimation bias of the SIV method. The mean values of the Relative Bias changed from −9.45% to −43.69% (−0.0088 to −0.0410 in Simple Bias) with a 34.24% increase in magnitude when d increased from 0.2 to 0.8. Their bias directions did not change, but the magnitudes became more prominent and the bias means were further away from zero. Similarly,

Noncomplier-Complier Level 2 Covariance Ratio also demonstrated similar influence, but its influence was much smaller. For each factor, the change in the bias means was almost proportional to the change of factor levels.

For the MMB-FC approach, although the mean values of the simple bias and the relative bias across all conditions were 0.0213 and 22.86%, showing overestimation, the mean values across different levels of several factors fluctuated greatly, both in their directions and magnitudes.

Noncomplier-Complier Level 2 Covariance Ratio demonstrated the most evident effect among the six factors. *Var* = 1 and 2 led to positive means (Simple = 0.0302/Relative = 33.38% and Simple = 0.1134/Relative = 123.02%), and *var* = 0.5 led to negative means (Simple = −0.0830/Relative = −91.92%). When compliers and non-compliers had the same Level 2 covariance matrix, the mean values were positive and minimum in their magnitudes. For *var* = 2 and 0, the mean value difference was 0.1964 in terms of the Simple Bias and 214.31 % in terms of the Relative Bias.

Complier Proportion had the second biggest influence, but the change mainly concentrated on changing *PC* from 0.1 to 0.3. With *PC* = 0.1, the mean values were positive and had big magnitudes, Simple = 0.0960/Relative = 103.53%. The numbers quickly became negative and trivial, Simple=−0.0039/Relative=−4.30%, when *PC* increased to 0.3. There was not much variation in the mean values among *PC* levels of 0.3, 0.5, and 0.8. The difference was 0.1005 between the biggest (when *PC* = 0.1) and the smallest (when *PC* = 0.5) mean values of the simple bias and 108.61 % between the two mean values of the relative bias. Note because the interaction term of

var and *PC* had a small-sized effect, the effect of one factor on the estimation bias would change across different levels of the other factor. The interaction effect is discussed more below.

Sample Size also had a considerable influence. With a small *n*, the mean values were sizable and positive. As *n* increased, the means became closer to zero. With $n = 50$, the mean values were 0.0684 for the Simple Bias and 74.89% for the Relative Bias, and with $n = 500$, they decreased to 0.0028 and 2.92%. However, as *n* kept increasing to 1,000, the mean values became negative with larger magnitudes, reaching −0.0111 and −12.23%. The difference was 0.0795 between the biggest (when $n = 50$) and the smallest (when $n = 1{,}000$) mean values of the simple bias and 87.12 % between the two mean values of the relative bias.

Mean Distance and Measurement Reliability demonstrated much smaller effects on estimation bias. The former had a minor negative effect, and the latter had a slight positive effect.

Table 11

Simple and Relative Estimation Bias Means by Different Configurations of PC and Var for the SIV and the MMB_FC Approaches

		Standard IV		MMB Full Constraint	
PC	**var**	**Simple Bias**	**Relative Bias**	**Simple Bias**	**Relative Bias**
0.1	0.5	**-0.0417**	**-46.68**	**-0.1340**	**-149.00**
	1.0	**-0.0544**	**-58.02**	**0.1137**	**125.31**
	2.0	**-0.0859**	**-87.99**	**0.3020**	**327.58**
0.3	0.5	**-0.0247**	**-26.96**	**-0.1384**	**-150.72**
	1.0	**-0.0308**	**-34.18**	**0.0120**	**13.89**
	2.0	**-0.0402**	**-43.84**	**0.1067**	**115.53**
0.5	0.5	**-0.0057**	**-6.10**	**-0.0553**	**-60.22**
	1.0	**-0.0068**	**-8.38**	**-0.0020**	**-2.73**
	2.0	**-0.0099**	**-10.44**	**0.0416**	**45.34**
0.8	0.5	**-0.0013**	**-1.97**	**-0.0069**	**-8.17**
	1.0	**-0.0017**	**-1.60**	**-0.0022**	**-2.07**
	2.0	**-0.0025**	**-2.68**	**0.0035**	**4.02**

Note. In the same row and within the same bias measure, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

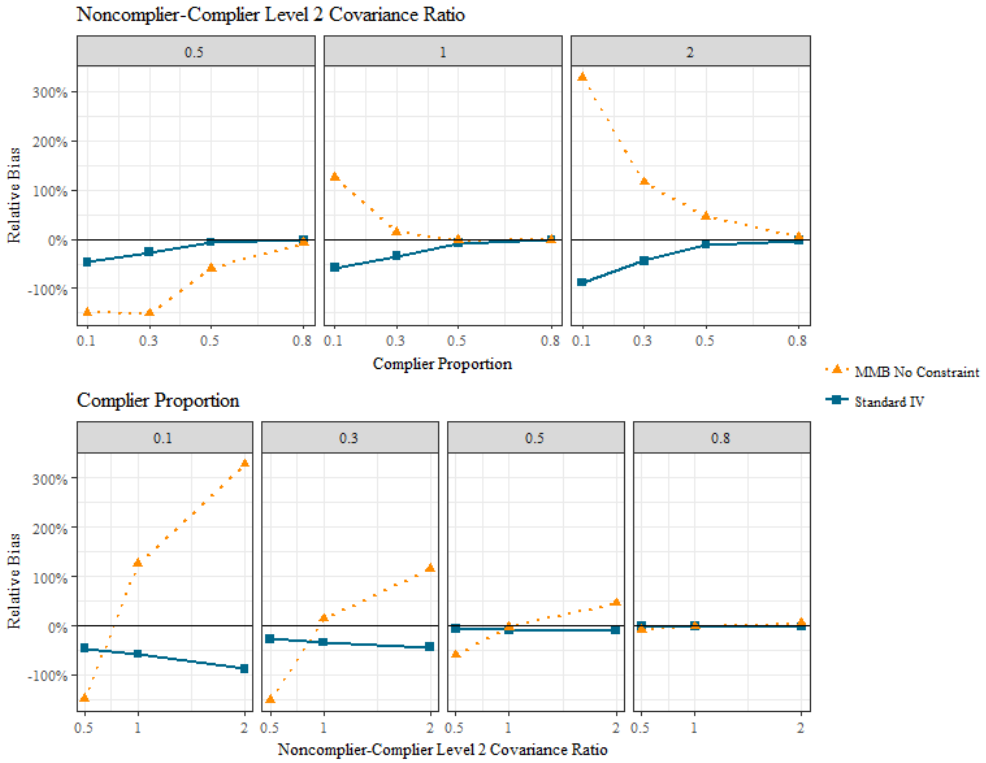

Figure 15. Simple and Relative Estimation Bias means by different configurations of *PC* and *var* for the SIV and the MMB_FC approaches

As discussed earlier, for both estimation bias measures, most partial η^2s obtained from the factorial ANOVA analyses were particularly small; therefore, an investigation of the only interaction term, "Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio", that met the 2% criterion was conducted. Table 11 summarizes the mean values of the simple and relative estimation bias by different configurations of *PC* and *var* for the SIV and the MMB_FC approaches. Figure 15 displays the interaction of the factors with two rows, and each row uses one factor as the X-axis variable and the other factor as the controlling variable. For example, the top row plots the change of the mean values of the relative bias over the four levels of the *PC* factor across the three levels of the *var* factor. Each plot in this row represents

one level of *var*. The change of the lines across the three plots in this row represents the influence of *PC* over different levels of *var*. Similarly, in the bottom row, the change of the lines across the four plots characterizes the influence of *var* on the relative bias over different levels of *PC*. The pattern of the simple bias measure was similar to that of the relative bias and was hence not presented.

In this figure, the lines for the SIV method changed slightly across different plots within a single row, suggesting the same conclusion from the factorial ANOVA analyses—the interaction effect between the two factors was small for the SIV method. In the main effect figure, as complier proportion became bigger or covariance ratio became smaller, the mean values of both bias measures for the SIV method decreased in their magnitudes and remained to be negative in their directions. After taking the interaction effect into consideration, the general trends of the two main effects did not shift much. The only observation worth mentioning was that the effect of *PC* was more prominent with *var* = 2 and the effect of *var* manifested itself more with *PC* = 0.1. The combination of a high *var* value and a low *PC* value led to more negative bias. With *var* = 2 and *PC* = 0.1, the mean values of the two bias measures reached the lowest, −0.0859 for the simple bias and −87.99% for the relative bias.

The lines for the MMB-FC method had a more considerable change across different plots within a single row. In the main effect figure, only when *PC* = 0.1, the mean values of the two bias measures were positive (Simple = 0.0960/Relative = 103.53%) and were all negative and very close to 0 at other *PC* levels. In a word, the main effect of the *PC* factor was not very clear. However, when controlling for the

var factor, the pattern became more evident. Within the same *var* level, the effect of *PC* could be summarized as "bigger complier proportion led to smaller bias magnitude, but different *PC* values were associated with different bias directions." The directions of the mean values were mainly determined by *var*. Within a fixed *PC* level, the effect of *var* was similar to its main effect: when *var* = 0.5, the estimation bias was negative on average; when *var* = 1, the average estimation bias became much closer to 0; when *var* = 2, the mean estimaiton bias was positive with the largest magnitude among the three *var* levels.

In conclusion, the combination of a lower *PC* and *var* = 0.5 led to a sizable negative bias, and the combination of a lower *PC* and *var* = 2 led to an especially large positive bias. With *var* = 0.5, the mean values were negative and had big magnitudes when *PC* = 0.1 (Simple = −0.1340/Relative = −149%) and *PC* = 0.3 (Simple = −0.1384/Relative = −150.72%). With *var* = 2 and *PC* = 0.1, the mean values were positive with even larger magnitudes (Simple = 0.3020/Relative = 327.58%). The smallest magnitude was obtained with *var* = 1 and *PC* = 0.8 (Simple = −0.0022/Relative = −2.07%).

Comparing the two estimation methods, the MMB-FC method on average yielded positive and lower magnitude bias, 0.0213 for Simple Bias and 22.86% for Relative Bias, while the SIV approach overall produced negative and bigger magnitude bias, −0.0252 and −27.14%. However, the MMB-FC method was more susceptible to adverse conditions. When *n* was low (smaller than 200) or *PC* was low (0.1), the estimation bias using the MMB-FC method on average surged on a much

larger scale than the SIV method, but with more favorable sample size (bigger than 200) or complier proportion (bigger than 0.1), the bias of the MMB-FC method on average was closer to zero.

In addition, the MMB-FC method also reacted more to the *var* factor. Unequal noncomplier- complier Level 2 covariance matrices exacerbated the estimation of the MMB-FC method on a large scale, especially when noncompliers had larger Level 2 covariance matrix. With *var* = 0.5, the estimation bias was on average negative, and with *var* = 2, positive. The former had a smaller magnitude than the latter. Factor *var*, on the other hand, was not that influential for the SIV approach. The *var* = 0.5 condition even yielded the smallest average estimation bias value for the SIV approach. Effect Size, Measurement Reliability, and Mean Distance had a similar influence on the magnitudes of the estimation bias means for both estimation methods. Only the directions were different.

Moreover, although the interaction effect of *PC***var* did not meet the 6% criterion for the MMB-FC method, it was larger than 2%. As a result, for this estimation method, the combination of a low *PC* and *var* = 0.5 led to a sizable negative bias; the combination of a low *PC* and *var* = 2 led to an especially large positive bias. For the SIV method, the interaction effect was not evident.

Results of factor effects with applicable conditions. This section examines the two estimation bias measures for the MMB-NC approach using only conditions that has more than 300 successfully estimated datasets per design cell. In other words,

Sample Size is bigger than or equal to 500, and Complier Proportion is bigger than or equal to 0.3.

Table 12

Factorial ANOVA Result: Estimation Bias Measures with the MMB-NC Method

		Max Partial Eta Square	
Estimation approach	**Term**	**Simple**	**Relative**
MMB No Constraint	Noncomplier-Complier Level 2 Covariance Ratio	0.02	0.02
	Sample Size*Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	0.02	NA
	Complier Proportion	0.01	NA
	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	NA	0.02
	Effect Size	NA	0.02

Note. $\eta^2 < 6\%$: grayed out. *NA* suggested that the term was not selected or was not one of the top three terms for the dependent variable.

Table 12 is a summary of the factorial ANOVA analysis. None of the terms was bigger than or equal to 6%. The biggest partial η^2 was only 0.02% for both the simple and relative bias measures.

Similar to the section for the other two estimation methods using all conditions, Table 13 summarizes the mean values of the simple and relative bias measures over different levels of the six factors, and Figure 16 and Figure 17 plots these values for the two dependent variables separately. Note that comparisons across the three methods were included after the examination of the MMB-NC approach, so Table 13, Figure 16, and Figure 17 also includes the results of the other two estimation methods using only sample units under applicable conditions.

Table 13

Simple and Relative Estimation Bias Means at each Level of the Factors for the Three Estimation Approaches under Applicable Conditions

		Standard IV		MMB No Constraint		MMB Full Constraint	
		Simple Bias	Relative Bias	Simple Bias	Relative Bias	Simple Bias	Relative Bias
Overall	Overall	-0.0023	-2.70	**-0.0015**	**-1.69**	**-0.0088**	**-9.63**
Sample Size	500	-0.0030	-3.66	**-0.0017**	**-2.09**	**-0.0087**	**-9.66**
	1000	-0.0017	-1.74	**-0.0013**	**-1.37**	**-0.0090**	**-9.61**
Complier Proportion	0.3	-0.0055	-5.97	**-0.0024**	**-2.41**	**-0.0240**	**-25.80**
	0.5	-0.0013	-1.82	**-0.0012**	**-1.51**	**-0.0027**	**-3.21**
	0.8	**-0.0002**	-0.29	**-0.0007**	**-1.04**	**0.0002**	**0.07**
Effect Size	0	-0.0025	NA	**-0.0015**	NA	**-0.0089**	NA
	0.2	-0.0028	-5.52	**-0.0017**	**-3.27**	**-0.0090**	**-17.72**
	0.5	-0.0019	-1.48	**-0.0014**	**-1.12**	**-0.0086**	**-6.81**
	0.8	-0.0022	-1.08	**-0.0013**	**-0.67**	**-0.0088**	**-4.37**
Measurement Reliability	0.5	-0.0026	-3.07	**-0.0017**	**-1.86**	**-0.0100**	**-10.84**
	0.8	-0.0020	-2.32	**-0.0013**	**-1.58**	**-0.0077**	**-8.43**
Mean Distance	0.2	-0.0011	-1.58	**-0.0010**	**-1.15**	**-0.0044**	**-5.12**
	0.5	-0.0020	-2.36	**-0.0015**	**-1.73**	**-0.0097**	**-10.57**
	0.8	-0.0038	-4.15	**-0.0020**	**-2.16**	**-0.0124**	**-13.21**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	**-0.0017**	**-2.01**	-0.0026	-2.70	**-0.0657**	**-71.32**
	1	**-0.0022**	**-2.33**	**-0.0001**	**0.00**	-0.0005	-0.39
	2	-0.0030	-3.74	**-0.0017**	**-2.27**	**0.0394**	**42.57**

Note. In the same row and within the same bias measure, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

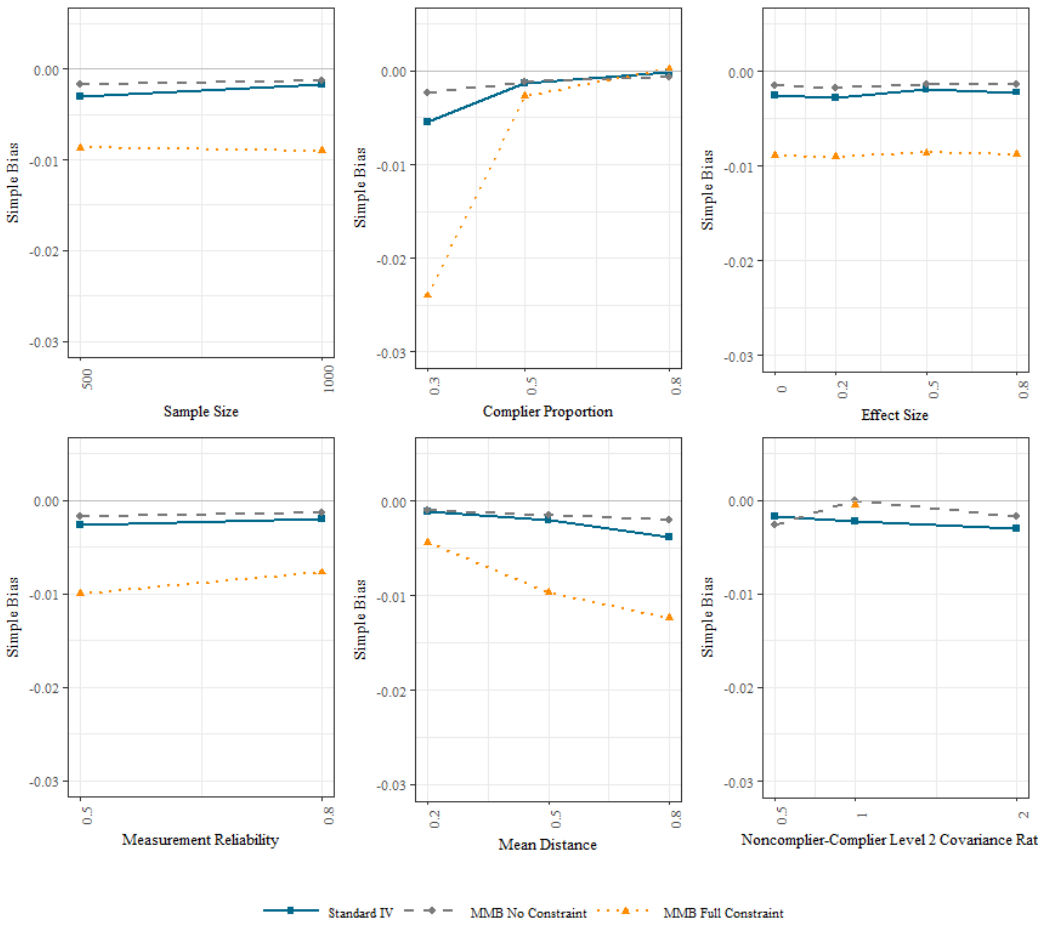

Figure 16. Simple Estimation Bias means at each level of the factors for the three estimation approaches under Applicable Conditions.

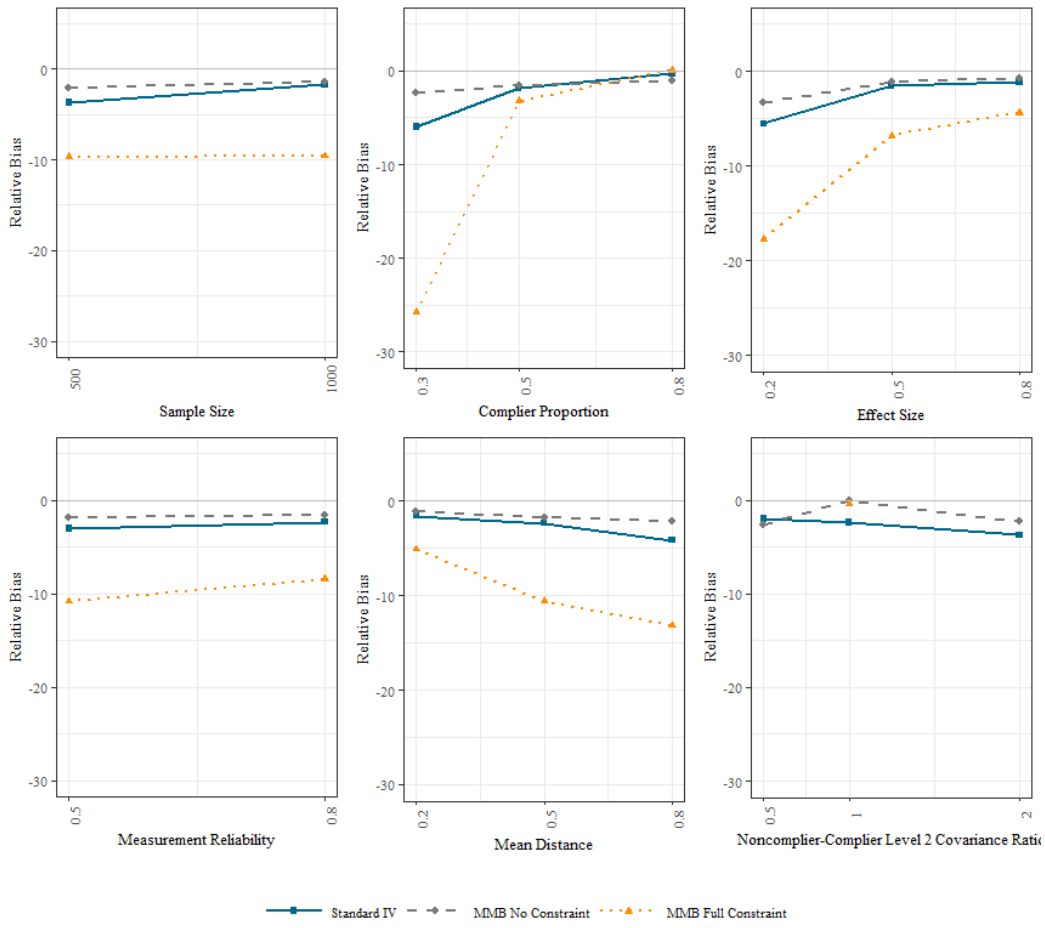

Figure 17. Relative Estimation Bias means at each level of the factors for the three estimation approaches under Applicable Conditions.

Because some *n* levels and *PC* levels that were very likely to lead to more extreme estimation bias were excluded from the analyses in this section, most mean values across different factor levels became much closer to zero, and there was less variation across different levels of a factor, especially for the SIV and the MMB-NC methods. To better display the trend for the MMB-NC method (although very small), the vertical axil limitations are set to be −0.03 to 0.005 for Figure 16 and −30% to 5% for Figure 17. For the MMB-FC estimation method, some mean values of the three

levels of *var* were out of these ranges, so in the "Noncomplier-Complier Level 2 Covariance Ratio" plot of each figure, the two mean values at the *var* = 0.5 and 1 levels could not be displayed.

Similar to the conclusion of the All Conditions section, Effect Size seemed to have different effects on the simple and the relative bias measures. However, this difference was mainly because the relative bias used the true treatment effect as the denominator. Therefore, Effect Size in fact did not show any effect on estimation bias for the MMB-NC method either.

Overall, the MMB-NC approach slightly underestimated the treatment effect (−0.0015/−1.69%). All main effects were extremely small, too. Factor *var* had the largest effect, where *var* = 1 led to the smallest bias on average. Complier Proportion had the second greatest effect. The influence of Sample Size, Measurement Reliability, and Mean Distance were negligible.

Comparing the three estimation methods under the Applicable Conditions, all three methods on average yielded negative bias. The MMB-NC method had the smallest average bias magnitude, −0.0015/−1.69%, the SIV approach had a slightly more sizable mean bias, −0.0023/−2.7%, and the MMB-FC method had a much more substantial average bias magnitude, −0.0088/−9.63%. For most sublevels within each factor, the MMB-NC method also performed the best, and the SIV method was not too much biased. The MMB-FC approach, on the other hand, performed much worse than the other two.

As shown in Table 13, most numbers in the two columns under the "MMB-NC" header were bold and underlined, indicating having the lowest absolute value among the three methods. Most numbers under the "MMB Full Constraint" header were just bold, indicating having the highest absolute value among the three methods. There are three exceptions. The first one was when *PC* = 0.8. The SIV and the MMB-FC approaches both had lower absolute simple bias means, and the MMB-NC approach had the largest absolute simple bias mean among the three. In terms of the relative bias means, the MMB-NC approach still had the biggest magnitude, but the MMB-FC approach had the smallest magnitude. The second exception was when *var* = 0.5: the SIV method, instead of the MMB-NC method, had the smallest bias magnitude for both simple and relative bias means. The last exception was when *var* = 1: the SIV method, instead of the MMB-FC method, had the largest bias magnitude for both simple and relative bias means.

Guidance on choosing n and rel with respect to parameter relative bias.

Table 14

Percentages of Cells with Cell Average Relative Estimation Bias within the Negligible Bounds[a]

		Standard IV				MMB No Constraint				MMB Full Constraint			
n	**rel**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**
50	0.5	11.11	0.00	18.52	51.85	NA	NA	NA	NA	0.00	7.41	11.11	48.15
50	0.8	22.22	0.00	33.33	77.78	NA	NA	NA	NA	0.00	0.00	22.22	59.26
100	0.5	14.81	3.70	51.85	81.48	NA	NA	NA	NA	3.70	3.70	29.63	77.78
100	0.8	11.11	11.11	62.96	92.59	NA	NA	NA	NA	11.11	7.41	25.93	81.48
200	0.5	3.70	33.33	74.07	***100.00***	NA	NA	NA	NA	0.00	18.52	29.63	85.19
200	0.8	11.11	25.93	81.48	***100.00***	NA	NA	NA	NA	0.00	33.33	33.33	88.89
500	0.5	3.70	70.37	92.59	***100.00***	NA	81.48	96.30	96.30	0.00	33.33	29.63	88.89
500	0.8	7.41	77.78	92.59	***100.00***	NA	77.78	96.30	***100.00***	3.70	29.63	40.74	***100.00***
1000	0.5	7.41	81.48	96.30	***100.00***	NA	92.59	***100.00***	***100.00***	3.70	33.33	33.33	***100.00***
1000	0.8	14.81	96.30	***100.00***	***100.00***	NA	96.30	***100.00***	***100.00***	7.41	33.33	40.74	***100.00***

Note. Percentages equal to 100% were italicized, bold, and underlined. *NA* were conditions excluded for the MMB-NC method.

[a]"Negligible Bounds" meant that the cell average Relative Estimation Bias was smaller than or equal to 10% and bigger than or equal to −10%.

As none of the factors met the 6% criterion, the recommendations on *n* and *rel* with respect to parameter estimation bias would not be conditional on the other four factors. The Complier Proportion factor was added because usable cells for the MMB-NC method conditioned on this variable. Therefore, Table 14 summarizes the 40 configurations using Sample Size, Measurement Reliability, and Complier Proportion. Within each configuration, there are 27 cells (3[*d*]*3[*var*]*3[*md*]). Each number in Table 14 represents the percentage of cells having cell relative bias mean value within the negligible bounds, −10% to 10%. All numbers equaling 100, indicating all cells with that configuration have mean values of their relative bias within the ± 10%

bounds, are italicized, bold, and underlined. Note that for the MMB-NC method, configurations excluded from the analysis are labeled with NA.

For all three methods, the number of configurations that had nested cells all meeting the negligible criterion was very low. Most configurations meeting the criterion concentrated on design cells with large complier proportion, 0.5 or 0.8, mostly 0.8. When $PC = 0.8$, to guarantee all design cells having negligible mean values of estimation bias for the SIV approach, n could be as small as 200 with or without highly reliable measurements (i.e., $rel = 0.8$). However, the requirement became much more restricted when $PC = 0.5$. Only the most optimistic sample size and reliability combinations would guarantee that all cells had negligible means. The MMB-NC method worked with $n = 500$ and $rel = 0.8$ if $PC = 0.8$. With $PC = 0.5$, only sample size of 1,000 would be adequate, while *rel* was irrelevant. Lastly, for the MMB-FC method, only $PC = 0.8$ was eligible for further consideration. The least restricted requirement was with $n = 500$ and $rel = 0.8$. When $n = 1{,}000$, there was no requirement on *rel*.

4.3.2. Standard error estimation bias

Results of factor effects with all conditions.

Table 15

Factorial ANOVA Result: SE Bias Measures with the SIV and the MMB-FC Methods

		Max Partial Eta Square	
Estimation approach	**Term**	**Simple**	**Relative**
Standard IV	Sample Size*Complier Proportion	**<u>21.48</u>**	**7.26**
	Sample Size	**6.75**	0.36
	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	2.20	NA
	Mean Distance*Sample Size*Complier Proportion	NA	0.32
MMB Full Constraint	Sample Size	**9.92**	5.04
	Sample Size*Complier Proportion	**7.46**	2.30
	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	1.27	NA
	Noncomplier-Complier Level 2 Covariance Ratio	NA	0.37

Note. $\eta^2 \geq 14\%$: bold and underlined ; $6\% \leq \eta^2 < 14\%$: bold; $\eta^2 < 6\%$: grayed out. *NA* suggested that the term was not selected or was not one of the top three terms for the dependent variable.

Table 15 is a summary of the linear regression results using the Simple SE Bias and the Relative SE Bias as dependent variables. In terms of the Simple SE Bias variable, for both the SIV approach and the MMB-FC approach, the "Sample Size" term and the "Sample Size*Complier Proportion" term met the 6% criterion and were therefore kept. The "Sample Size*Complier Proportion" term was the most influential term for the SIV approach, accounting for 21.48% of the sum of variance of that term plus error variance. This term had a "large" effect size. "Sample Size" had the second largest partial η^2, 6.75%, for the SIV approach, meaning that this term explained 6.75% of the sum of variance of that term plus error variance. The MMB-FC approach had the same two terms kept, but the Sample Size term had a larger partial

η^2, 9.92%, and the interaction term of *n* and *PC* had the second largest partial η^2, 7.46%. In terms of the Relative SE Bias variable, on the other hand, only the "Sample Size*Complier Proportion" term was kept for the SIV approach (partial η^2 = 7.26%). No term was kept for the MMB-FC approach.

As one interaction term was kept, the following discussion presents the main effect results first and then the interaction effect.

Table 16

Simple and Relative SE Bias Means at each Level of the Factors for the SIV and the MMB-FC Approaches

		Standard IV		MMB Full Constraint	
		Simple SE Bias	Relative SE Bias	Simple SE Bias	Relative SE Bias
Overall	Overall	-0.0548	-1.33	-0.0826	-15.17
Sample Size	50	-0.1229	-3.75	-0.1897	-34.68
	100	-0.0857	-1.03	-0.1172	-20.97
	200	-0.0590	-1.39	-0.0793	-13.53
	500	-0.0279	-2.70	-0.0459	-8.61
	1000	0.0036	1.70	-0.0232	-5.61
Complier Proportion	0.1	-0.2474	-20.08	-0.2530	-41.96
	0.3	0.0008	4.29	-0.0633	-14.23
	0.5	0.0187	8.45	-0.0146	-4.25
	0.8	0.0016	1.21	-0.0010	-0.44
Effect Size	0	-0.0587	-1.38	-0.0804	-14.90
	0.2	-0.0574	-1.80	-0.0833	-15.10
	0.5	-0.0539	-1.03	-0.0835	-15.31
	0.8	-0.0491	-1.12	-0.0833	-15.34
Measurement Reliability	0.5	-0.0594	-0.95	-0.0880	-13.18
	0.8	-0.0505	-1.68	-0.0781	-16.83
Mean Distance	0.2	-0.0580	-1.09	-0.0917	-16.61
	0.5	-0.0498	-1.09	-0.0832	-15.46
	0.8	-0.0565	-1.81	-0.0730	-13.44
Noncomplier-Complier Level 2 Covariance Ratio	0.5	-0.0412	-1.85	-0.0886	-17.90
	1	-0.0522	-1.24	-0.0615	-11.89
	2	-0.0709	-0.90	-0.0979	-15.78

Note. In the same row and within the same bias measure, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

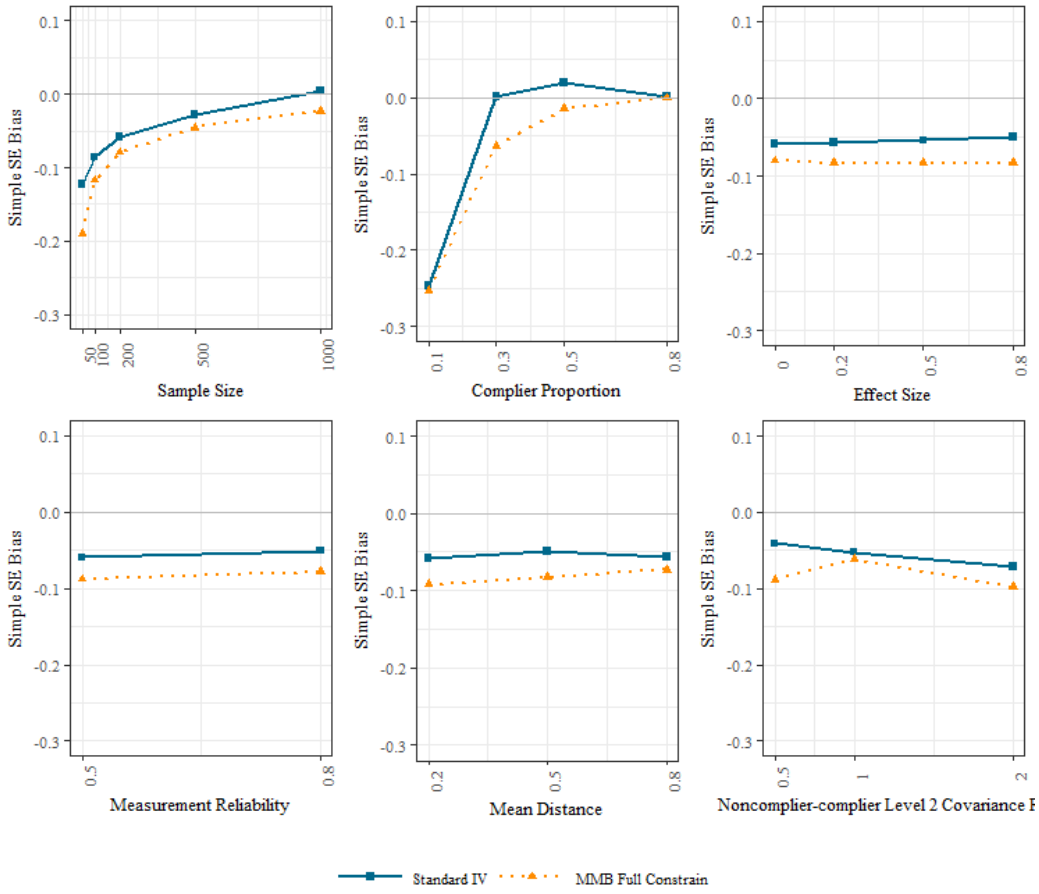

Figure 18. Simple SE Bias means at each level of the factors for the SIV and the MMB_FC approaches.

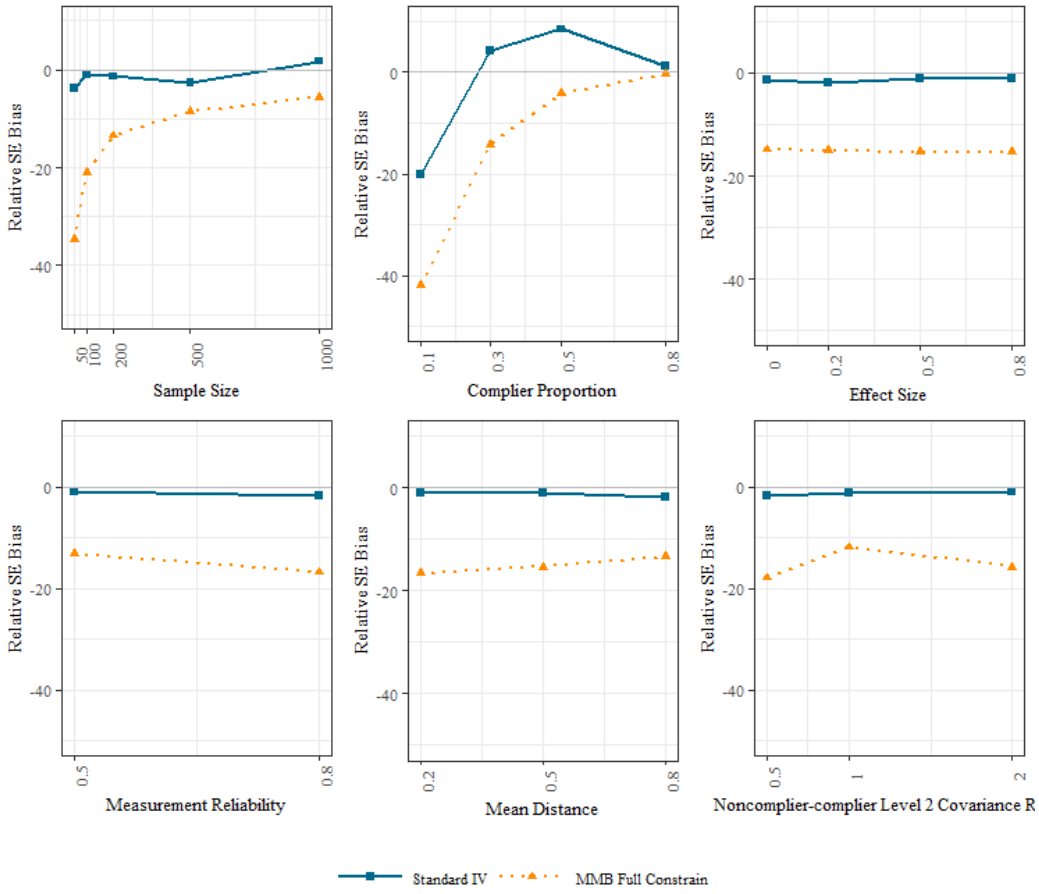

Figure 19. Relative SE Bias means at each level of the factors for the SIV and the MMB_FC approaches.

Table 16 summarizes the mean values of the simple SE bias and the relative SE bias over different levels of the six factors, and Figure 18 and Figure 19 plots these values for the two dependent variables separately. The SIV approach on average slightly underestimated the SE, resulting in an overall Simple SE Bias of −0.0548 and Relative SE Bias of −1.33%. The MMB-FC approach on average also underestimated the SE, but on a larger scale, especially for the Relative Bias. The overall Simple SE Bias was −0.0826, less than two times of that of the SIV approach, but the overall Relative SE Bias reached −15.17%, 11 times of that of the SIV approach.

The trends of the two bias measures over different levels of the six factors differed between the two bias measures. Therefore, the results of Simple SE Bias are discussed first, and the results of Relative SE Bias are next.

In the matter of Simple SE Bias, *PC* and *n* had the most evident effects for both estimation approaches. The effect of *PC* was completely positive for the MMB-FC approach but was partially positive for the SIV method. With the MMB-FC approach, the mean values of the simple SE bias increased together with *PC* and became the closest to zero with *PC* = 0.8. When *PC* = 0.1, the simple bias mean was −0.2530, but it quickly changed to −0.0633, when *PC* turned to 0.3. With *PC* = 0.8, the mean almost diminished to zero, −0.0010. The decrease in the magnitude of the mean values was 0.2520. The mean values for the SIV approach, by contrast, started with the lowest value, −0.2474, with *PC* = 0.1, culminated at 0.0187 when *PC* = 0.5, and decreased to 0.0016 when *PC* continued to change from 0.5 to 0.8. When *PC* = 0.3 or 0.8, the mean values had smaller magnitudes. The difference between the largest and the smallest means was 0.2661, slightly bigger than that of the MMB-FC method.

With bigger Sample Size, both the SIV method and the MMB-FC method had higher Simple SE Bias on average. The simple SE bias mean increased from −0.1229 to 0.0036 with a 0.1265 increase when changing the sample size from 50 to 1,000 using the SIV approach, and the change was from −0.1897 to −0.0232 with an increase of 0.1665 for the MMB-FC approach. For the SIV method, when *n* = 1,000, the mean turned to be positive while the means of all other *n* levels were negative. Its

magnitude, however, was still the smallest. As *PC* and *n* also had a sizable interaction effect, the effect of one factor might change across different levels of the other factor.

The other four factors did not display as evident effects as the two factors mentioned above. The *var* factor had the third greatest effect for both estimation methods, but the effect was much smaller than *PC* and *n.* Across all levels of the *var* factor, the means were all negative for both methods. For the SIV approach, the means became more negative as *var* increased, and the bias magnitude was the greatest when *var* = 2. For the MMB-FC method, *var* = 1 led to the least negative bias mean, i.e., the smallest magnitude mean among the three *var* levels.

The Effect Size factor, the Measurement Reliability factor and the Mean Distance factor had the fourth to sixth greatest impact on the Simple SE Bias for the SIV approach. For the MMB-FC approach, the Mean Distance factor, the Measurement Reliability, and the Effect Size factor ranked the fourth to sixth in terms of their impacts. For both estimation methods, the impacts of these three factors were very small.

With regard to Relative SE Bias, *PC* and *n* also had the most evident effects for both estimation approaches. The trend patterns were also similar to those of Simple SE Bias. For the MMB-FC approach, *PC* had a completely positive effect on Relative SE Bias: the mean value of relative bias increased from −41.96% to −0.44% with *PC* changing from 0.1 to 0.8, resulting in an increase of 41.52%.

For the SIV method, the effect of *PC* was partially positive. When *PC* increased from 0.1 to 0.5, the mean value of Relative Bias increased from −20.08% to

8.45%, but decreased to 1.21% when *PC* = 0.8. The highest mean value of bias was at *PC* = 0.5, but when *PC* = 0.8, the value was the closest to zero. The difference between the largest and the smallest mean values was 28.53%, much smaller than that of the MMB-FC method.

For the MMB-FC method, the effect of the *n* factor on Relative SE Bias was similar to its effect on Simple SE Bias. With bigger *n*, the mean value of Relative SE Bias was higher. The value increased from −34.68% to −5.61% (29.07% difference) when *n* changed from 50 to 1,000. However, the effect of *n* did not have a clear pattern for the SIV approach. The biggest mean value for this estimation method was 1.7% when *n* = 1,000, and the smallest was −3.75% when *n* = 50. Nonetheless, when *n* = 100, the mean value of bias was the closest to zero, −1.03%. The difference between the biggest and the smallest means across the five *n* levels was only 5.45%. The reason that the *n* factor exemplified different effects on the two bias measures was because that *n* had effect on the empirical SE. In the current case, the empirical SE became smaller with bigger sample size. Because the interaction term of *n* and *PC* also had a sizable effect, the effects of the two factors could also change across different levels of the other factor.

The other four factors did not display as evident effects as *n* and *PC* for the Relative SE Bias measure either. Again, factor *var*, following *n* and *PC*, exhibited the third greatest effect for both estimation methods. The effect was positive but negligible for the SIV approach. As for the MMB-FC method, the effect was slightly

larger, and when *var* = 1, the mean value of bias was the highest and closest to zero among the three *var* levels.

The other three factors had minimum influences on the two estimation methods, especially for the SIV method. The Effect Size factor, the Measurement Reliability factor, and the Mean Distance factor had the fourth to sixth greatest impact on the Relative SE Bias for the SIV approach. For the MMB-FC approach, *rel*, *md*, and *d* were the three factors having the fourth to sixth highest impact on the Relative SE Bias.

Table 17

Simple and Relative SE Bias Means by Different Configurations of PC and N for the SIV and the MMB_FC Approaches

		Standard IV		MMB Full Constraint	
n	**PC**	**Simple SE Bias**	**Relative SE Bias**	**Simple SE Bias**	**Relative SE Bias**
50	0.1	**-0.4976**	**-33.97**	**-0.4080**	**-58.26**
	0.3	**-0.0888**	**-10.58**	**-0.2300**	**-44.14**
	0.5	**0.0673**	**22.83**	**-0.0878**	**-26.50**
	0.8	**0.0074**	**4.71**	**-0.0101**	**-6.39**
100	0.1	**-0.4391**	**-33.17**	**-0.3535**	**-52.13**
	0.3	**0.0425**	**11.14**	**-0.1141**	**-28.91**
	0.5	**0.0261**	**14.20**	**-0.0115**	**-5.11**
	0.8	**0.0013**	**1.20**	**0.0010**	**0.77**
200	0.1	**-0.2857**	**-26.54**	**-0.2911**	**-46.19**
	0.3	**0.0318**	**13.19**	**-0.0369**	**-13.03**
	0.5	**0.0064**	**5.17**	**0.0028**	**2.25**
	0.8	**0.0010**	**1.42**	**0.0012**	**1.67**
500	0.1	**-0.1192**	**-16.55**	**-0.1865**	**-36.32**
	0.3	**0.0053**	**3.90**	**0.0008**	**0.98**
	0.5	**0.0012**	**1.62**	**0.0005**	**0.63**
	0.8	**0.0000**	**0.05**	**0.0000**	**-0.04**
1000	0.1	**0.0122**	**3.97**	**-0.0948**	**-24.12**
	0.3	**0.0018**	**2.00**	**0.0015**	**1.67**
	0.5	**0.0006**	**1.12**	**0.0002**	**0.30**
	0.8	**-0.0001**	**-0.30**	**-0.0001**	**-0.35**

Note. In the same row and within the same bias measure, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

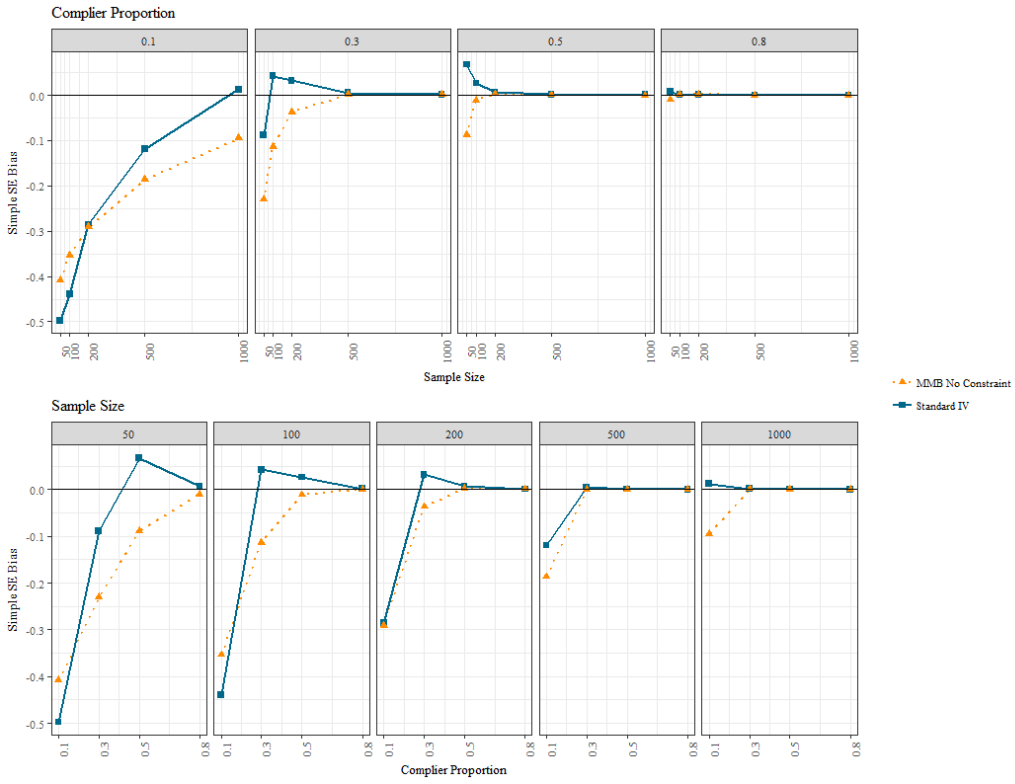

Figure 20. Simple SE Bias means by different configurations of *PC* and *n* for the SIV and the MMB_FC approaches

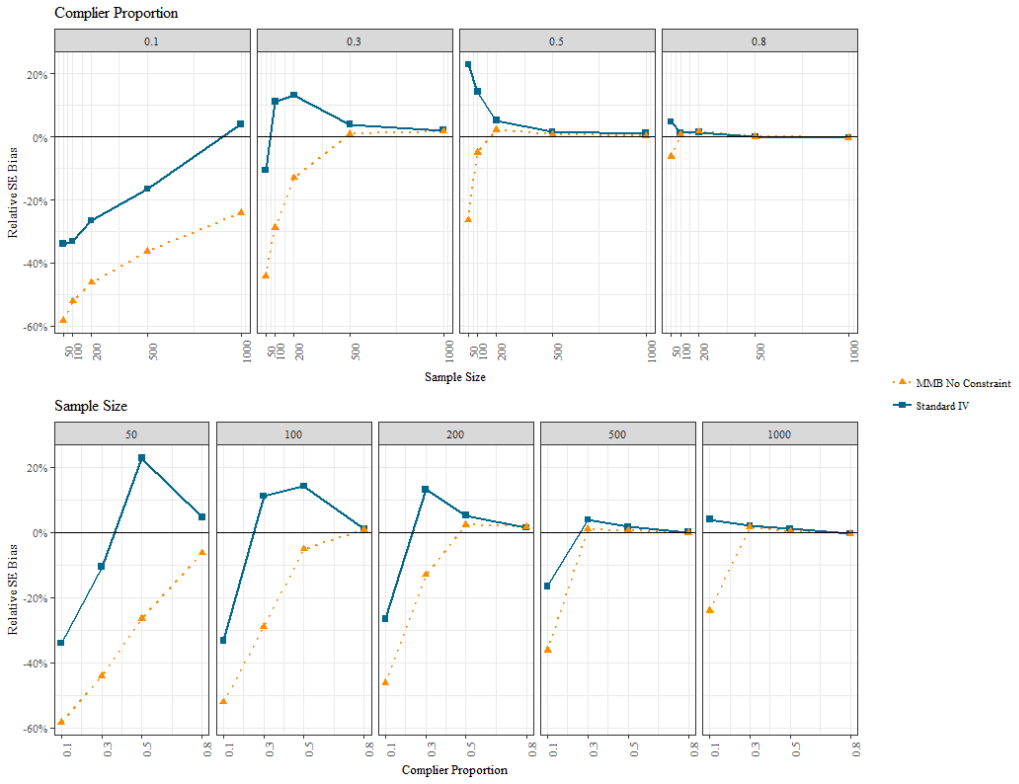

Figure 21. Relative SE Bias means by different configurations of *PC* and *n* for the SIV and the MMB_FC approaches

As mentioned earlier, the interaction term "Sample Size*Complier Proportion" was kept for the SIV approach using both simple and relative SE bias measures and was chosen for the MMB-FC approach using the simple SE bias measure. Table 17 summarizes the means for each unique configuration of *PC* and *n* for both simple and relative measures. Figure 20 and Figure 21 depict the interaction effect. The result of the SIV approach is discussed first below.

In Figure 20, the lines of the SIV method evidently changed across different plots within a single row. When only looking at the main effect of *n*, factor *n* displayed a positive effect: larger *n* was followed by larger mean value. Because the

mean values for all four levels (from n = 50 to 500) were negative, larger mean values indicated closer to zero values. When n = 1,000, the mean value was positive, but its magnitude was also the smallest and closest to zero. After accounting for *PC*, the impact of n changed considerably across different levels of the *PC* factor. With *PC* = 0.1, the trend line was similar to the trend line in the main effect plot, only with a more evident magnitude change: the mean value of the simple bias increased from −0.4976 to 0.0122 (0.5098 change in magnitude). However, with larger *PC* values, the trend lines became very different. With *PC* = 0.3, the mean started with a negative value, −0.0888, when n was 50, but it became positive and closer to zero, 0.0425, when n changed to 100. As n kept increasing, the mean remained positive and became even closer to zero. When *PC* = 0.5 and 0.8, the impact of n became negative: larger n was followed by smaller mean value, and because the mean values at all five levels were positive, the magnitudes were getting closer zero. The only exception was when n = 1,000 and *PC* = 0.8, and the mean value of Simple SE Bias was −0.0001. However, the magnitude was too small to be considered as an exception. A similar pattern can be observed for Relative Bias, in Figure 21. The only difference was when *PC* = 0.3, the maximum mean value occurred with sample size of 200 instead of 100.

The pattern change of the impact of *PC* on Simple SE Bias over different levels of n was displayed in the second row of Figure 20. Recall that the main effect of *PC* was positive first and then negative. The highest mean value occurred at *PC* = 0.5, but when *PC* = 0.3 or 0.8, the means were very close to zero. After taking n into account, the impact of *PC* did not change drastically across n levels of 50 to 500. The

shapes of the four trend lines were similar to the trend line of the main effect: starting at a very low negative value, increasing first to become positive, and then decreasing with the magnitude getting close to zero. When n = 1,000, the mean value was positive and close to zero, 0.0122, even with PC = 0.1. As PC became bigger, the mean value decreased and approximated zero. Relative Bias, in the second row of Figure 21, almost had the same pattern and is hence not discussed more here.

The general interaction effect of n and PC on the two bias measures for the SIV method could be summarized as this: low n and low PC yielded negative bias on average with especially large magnitude, and if increasing n or PC alone, the bias could become positive and eventually close to zero with a very favorable n level or PC level.

In terms of the MMB-FC method, the general interaction effect of the two factors on its SE bias, both simple and relative, were almost the same as the interaction effect for the SIV approach. The only difference was that the SE estimation resulting from the MMB-FC method was more conservative than the SIV method. In other words, under conditions that were more likely to lead to SE underestimation, the MMB-FC method would yield more underestimation than the SIV method. Even under conditions where the SIV method had SE overestimation, the MMB-FC method was still more likely to yield underestimation. When both methods were accurate about their SE estimations, the two methods performed very similarly. The only exception was when n = 50 or 100 and PC = 0.1, the MMB-FC method had slightly less underestimation than the SIV method, where the former had

mean values of simple bias at −0.4976 and −0.4391 and the latter had −0.4080 and −0.3535.

Comparing the two estimation methods, both methods yielded, on average, negative bias, but the MMB-FC approach had a larger magnitude. With respect to the same sublevel of a factor, the MMB-FC approach always had a smaller mean than the SIV approach. As shown in Table 16, when both the SIV and the MMB-FC approaches had negative means, the latter always had higher magnitude than the former. When the SIV approach had positive means, the MMB-FC method still had negative means. This was true for both simple bias and relative bias measures. In other words, the SE estimation of the MMB-FC method was more conservative than the SIV method, and the MMB-FC method would yield more underestimation than the SIV method.

Both estimation methods were mostly influenced by *n* and *PC*. With bigger *n* or *PC*, both estimation methods yielded closer to zero simple bias means and relative bias means. The difference between the two estimation methods only diminished with the combination of higher sample size and complier proportion.

Results of factor effects with applicable conditions.

Table 18

Factorial ANOVA Result: SE Bias Measures with the MMB-NC Method

		Max Partial Eta Square	
Estimation approach	**Term**	**Simple**	**Relative**
MMB No Constraint	Noncomplier-Complier Level 2 Covariance Ratio	0.55	0.68
	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	0.50	0.38
	Sample Size	0.18	0.12

Note. $\eta^2 < 6\%$: grayed out.

Table 18 displays the factorial ANOVA analysis results for the MMB-NC approach using only applicable conditions. None of the terms met the 6% criterion. The largest partial η^2, for the main effect of the *var* factor was only 0.55% with the Simple Bias measure and 0.68% with the Relative Bias measure. As no term was picked up, only the main effects were analyzed.

Table 19

Simple and Relative SE Bias Means at each Level of the Factors for the Three Approaches

		Standard IV		MMB No Constraint		MMB Full Constraint	
		Simple SE Bias	Relative SE Bias	Simple SE Bias	Relative SE Bias	Simple SE Bias	Relative SE Bias
Overall	Overall	0.0015	1.40	**0.0027**	**3.05**	**0.0005**	**0.53**
Sample Size	500	0.0022	1.86	**0.0039**	**3.98**	**0.0004**	**0.52**
	1000	0.0008	0.94	**0.0018**	**2.29**	**0.0005**	**0.54**
Complier Proportion	0.3	0.0036	2.95	**0.0060**	**5.82**	**0.0011**	**1.33**
	0.5	0.0009	1.37	**0.0015**	**2.44**	**0.0003**	**0.47**
	0.8	**0.0000**	**-0.12**	**0.0002**	**0.41**	-0.0001	-0.19
Effect Size	0	0.0013	1.20	**0.0027**	**2.97**	**0.0000**	**0.19**
	0.2	0.0014	1.13	**0.0026**	**2.78**	**0.0003**	**0.32**
	0.5	0.0019	1.87	**0.0029**	**3.48**	**0.0011**	**1.04**
	0.8	0.0013	1.40	**0.0026**	**2.95**	**0.0005**	**0.57**
Measurement Reliability	0.5	0.0016	1.38	**0.0038**	**3.77**	**0.0006**	**0.69**
	0.8	0.0013	1.42	**0.0020**	**2.58**	**0.0003**	**0.37**
Mean Distance	0.2	0.0012	1.26	**0.0031**	**3.62**	**0.0004**	**0.77**
	0.5	0.0014	1.23	**0.0026**	**2.82**	**0.0004**	**0.30**
	0.8	0.0018	1.70	**0.0024**	**2.73**	**0.0006**	**0.53**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	**0.0012**	**1.50**	**0.0002**	**0.29**	**0.0002**	0.59
	1	0.0014	1.30	**0.0044**	**4.71**	**0.0008**	**0.65**
	2	0.0018	1.40	**0.0036**	**4.22**	**0.0004**	**0.36**

Note. In the same row and within the same bias measure, the number with the biggest absolute value was bold, and the smallest absolute value was bold and underlined.

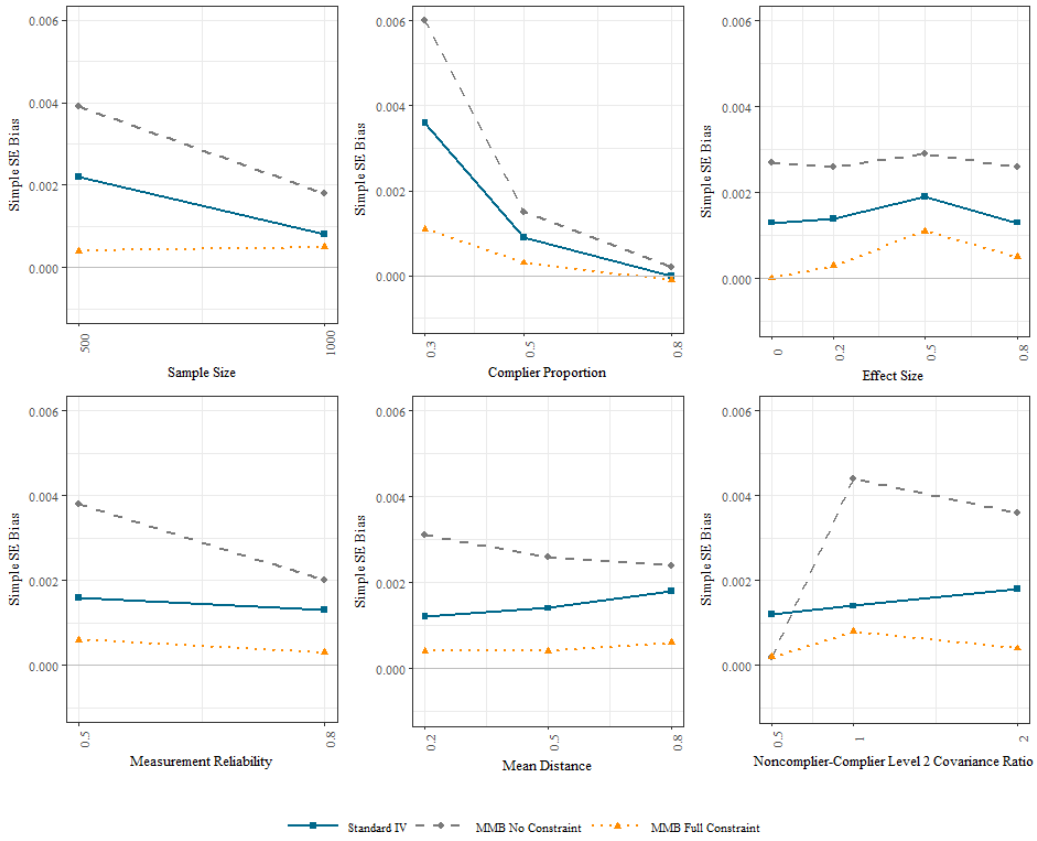

Figure 22. Simple SE Bias means at each level of the factors for the three approaches.

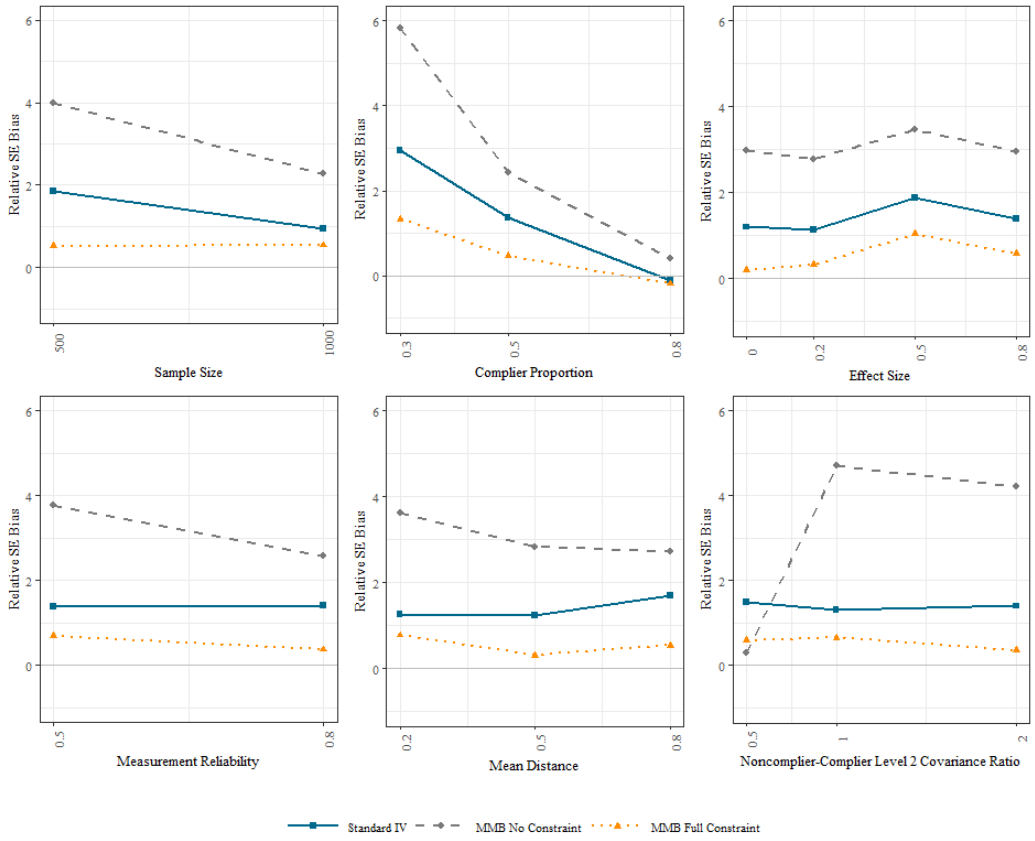

Figure 23. Relative SE Bias means at each level of the factors for the three approaches.

Table 19 summarizes the mean values of the simple SE bias and the relative SE bias over different levels of the six factors, and Figure 22 and Figure 23 plot these values for the two dependent variables respectively. Note that Figure 22 restricts the vertical axis from −0.001 to 0.006, and Figure 23 restricts from −1% to 6% to display the small variations across different levels within a factor. Overall, the MMB-NC method slightly overestimated the SE. The overall mean value of Simple SE Bias was 0.0027, and the mean value of Relative Bias was 3.05%. The variations among the mean values across different sublevels within one factor were extremely small. For

both bias measures, Complier Proportion had the greatest impact, and Effect Size had the smallest. As Simple Bias and Relative Bias behaved very similar for the MMB-NC approach, they are discussed together below.

The Sample Size factor, the Complier Proportion factor, the Measurement Reliability factor and the Mean Distance factor all had a negative effect on the SE bias: higher *n*, *PC*, *rel* or *md* value led to smaller and closer to zero bias means. There was no clear pattern for the Effect Size factor. While for the Covariance Ratio factor, *var* = 1 led to the highest and farthest-from-zero bias, and *var* = 0.5 led to the lowest and close-to-zero bias.

To summarize, under Applicable Conditions, the overall mean values for the MMB-NC method, 0.0027 and 3.05%, were the biggest and farthest from zero among the three estimation methods. The MMB-FC method had the smallest and closest to zero overall means. For most sublevels, the MMB-FC method also had the smallest and closest to zero means, and the MMB-NC method had the biggest and the farthest from zero means. The only exception was when *var* = 0.5, the SIV approach had the biggest mean and the MMB-NC method had the smallest. The differences among the three methods were very small, and they also became smaller with higher sample size, higher complier proportion, and higher measurement reliability.

Guidance on choosing n and rel with respect to standard error relative bias.

Table 20

Percentages of Cells with Cell Average Relative SE Bias Meeting the Negligible Criterion[a]

		Standard IV				MMB No Constraint				MMB Full Constraint			
n	**rel**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**
50	0.5	0.00	41.67	0.00	86.11	NA	NA	NA	NA	0.00	0.00	8.33	91.67
50	0.8	0.00	50.00	8.33	97.22	NA	NA	NA	NA	0.00	0.00	0.00	61.11
100	0.5	0.00	38.89	11.11	***100.00***	NA	NA	NA	NA	0.00	8.33	77.78	97.22
100	0.8	0.00	41.67	5.56	***100.00***	NA	NA	NA	NA	0.00	0.00	61.11	***100.00***
200	0.5	0.00	13.89	97.22	***100.00***	NA	NA	NA	NA	0.00	44.44	***100.00***	97.22
200	0.8	0.00	25.00	97.22	***100.00***	NA	NA	NA	NA	0.00	44.44	***100.00***	***100.00***
500	0.5	16.67	***100.00***	***100.00***	***100.00***	NA	52.78	97.22	***100.00***	0.00	***100.00***	***100.00***	***100.00***
500	0.8	11.11	97.22	***100.00***	***100.00***	NA	80.56	97.22	***100.00***	0.00	97.22	***100.00***	***100.00***
1000	0.5	58.33	***100.00***	***100.00***	***100.00***	NA	63.89	***100.00***	***100.00***	22.22	***100.00***	***100.00***	***100.00***
1000	0.8	61.11	***100.00***	***100.00***	***100.00***	NA	91.67	***100.00***	***100.00***	19.44	***100.00***	***100.00***	***100.00***

Note. Percentages equal to 100% were italicized, bold, and underlined. NA were conditions excluded for the MMB-NC method.

[a]"Negligible Criterion" meant that the cell average relative SE bias was within the bound of −10% and 10%.

Table 20 uses cross tabulation to summarize Relative SE Bias with respect to the two factors identified above (i.e., *n* and *PC*). Each number in this table is a summary of the 36 cells with the same configuration of *n*, *PC*, and *rel*. Each number represents the percentage of cells with cell average Relative SE Bias within the negligible bound, −10% to 10%. Numbers equaling 100% were italicized, bold, and underlined. Note that for the MMB-NC method, configurations excluded from the analyses were labeled with NA.

For the SIV approach, when *PC* was 0.1, none of the cells could yield an average relative SE bias within the negligible bounds. When *PC* = 0.3, Sample Size of 500 and Measurement Reliability of 0.5 could guarantee an average relative SE bias within the negligible bounds for all cells nested within. If *rel* was improved to 0.8, the percentage actually became lower a little bit, 97.22%. However, 97.22% was very close to 100%, it could be mainly due to random variation. When *PC* = 0.5 and 0.8, Sample Size of 500 and 100 would respectively guarantee an average relative SE bias within the negligible bounds for all cells nested within, irrespective of the *rel* values.

The MMB-NC approach had more limited conditions than the SIV approach. When *PC* = 0.3, none of the cells could yield an average relative SE bias within the negligible bounds. When *PC* = 0.5, only *n* = 1,000 could guarantee an average relative SE bias within the negligible bounds for all cells nested within. When *PC* = 0.8, *n* = 500 was enough.

For the MMB-FC approach, when *PC* was 0.1, none of the cells met the requirement. When *PC* = 0.3, Sample Size of 500 and Measurement Reliability of 0.5 could guarantee an average relative SE bias within the negligible bounds for all cells nested within. If *rel* was improved to 0.8, the percentage actually became lower a little bit, 97.22%. Again, 97.22% was very close to 100%, it could be mainly due to random variation. When *PC* = 0.5, *n* = 200 was enough. With *PC* = 0.8, *n* could be as low as 100 or 200 with *rel* = 0.8. Even with *rel* = 0.5 (*PC* = 0.8 and *n* = 100 or 200), 97.22% of the design cells nested within had mean Relative SE Bias within the negligible bounds. If *n* reached 500, *rel* became irrelevant.

4.4.Power and Type I Error

This section presents the results regarding research questions three and four: how will the six factors affect the statistical power and the type I error rate using the SIV and MMB methods? The Significant Indicator was the dependent variable used in this section.-Section 4.4.1 presents the analysis result of power, and section 4.4.2 type I error. Within each of these two sections, the analysis results for the SIV and the MMB-FC approaches using all conditions are discussed first, the analysis results for the MMB-NC approach using only "Applicable Conditions" follows, and the recommendations on choosing sample size and measurement reliability are presented in the end.

4.4.1. Power

Results of factor effects with all conditions. This section examined the empirical power for the SIV and the MMB-FC approaches using all simulation conditions.

Table 21

Factorial ANOVA Result: Power with the SIV and the MMB-FC Methods

		Max Pseudo Partial Eta Squares			
Estimation Approach	**Term**	**Sample 1**	**Sample 2**	**Sample 3**	***Mean of All Maxes***
Standard IV	Complier Proportion	**18.46**	**18.01**	**18.24**	***18.24***
	Sample Size	**14.60**	**13.93**	**13.97**	***14.17***
	Effect Size	**13.90**	**14.00**	**14.39**	***14.10***
MMB Full Constraint	Sample Size*Complier Proportion	**7.62**	**7.44**	**7.33**	*7.46*
	Effect Size	**7.17**	**7.35**	**7.47**	*7.33*
	Effect Size*Complier Proportion	3.80	3.93	3.88	*3.87*

Note. $\eta^2 \geq 14\%$: bold and underlined; $6\% \leq \eta^2 < 14\%$: bold; $\eta^2 < 6\%$: grayed out.

Table 21organizes the results of the factorial ANOVA analyses for the SIV and the MMB-FC estimation methods with sample unites under all conditions. For the SIV approach, three main effect terms met the 6% criterion and were therefore kept. In fact, all three terms had "Large" effects because the values of their "Mean of All Maxes" all exceeded the 14% cutoff line. The Complier Proportion term was the most influential among all terms, accounting for 18.24% of the sum of the deviance reduced by this term plus the residual deviance. Sample Size and Effect Size had the second and third largest pseudo partial η^2s, 14.17% and 14.10%. On the other hand, for the MMB-FC approach, only two terms met the 6% criterion. The term with the largest pseudo partial η^2 was the interaction term of $n*PC$. This term had a Pseudo partial η^2 of 7.46%. The main effect of factor d accounted for 7.33% of the sum of the deviance reduced by this term plus the residual deviance. The two terms both had a "Medium" sized pseudo partial η^2.

Table 22

Mean Values of Power at each Level of the Factors for the SIV and the MMB-FC Approaches

		Standard IV	MMB Full Constraint
Overall	Overall	**0.2576**	**0.3760**
Sample Size	50	**0.0946**	**0.3123**
	100	**0.1441**	**0.2812**
	200	**0.2180**	**0.3162**
	500	**0.3409**	**0.4192**
	1000	**0.4420**	**0.5106**
Complier Proportion	0.1	**0.0832**	**0.3505**
	0.3	**0.1591**	**0.2870**
	0.5	**0.2956**	**0.3535**
	0.8	**0.4911**	**0.5132**
Effect Size	0.2	**0.0988**	**0.2124**
	0.5	**0.2691**	**0.3893**
	0.8	**0.4052**	**0.5261**
Measurement Reliability	0.5	**0.2236**	**0.3380**
	0.8	**0.2886**	**0.4078**
Mean Distance	0.2	**0.2621**	**0.3649**
	0.5	**0.2596**	**0.3770**
	0.8	**0.2510**	**0.3860**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	**0.2943**	**0.3418**
	1	**0.2588**	**0.3603**
	2	**0.2196**	**0.4247**

Note. In the same row, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

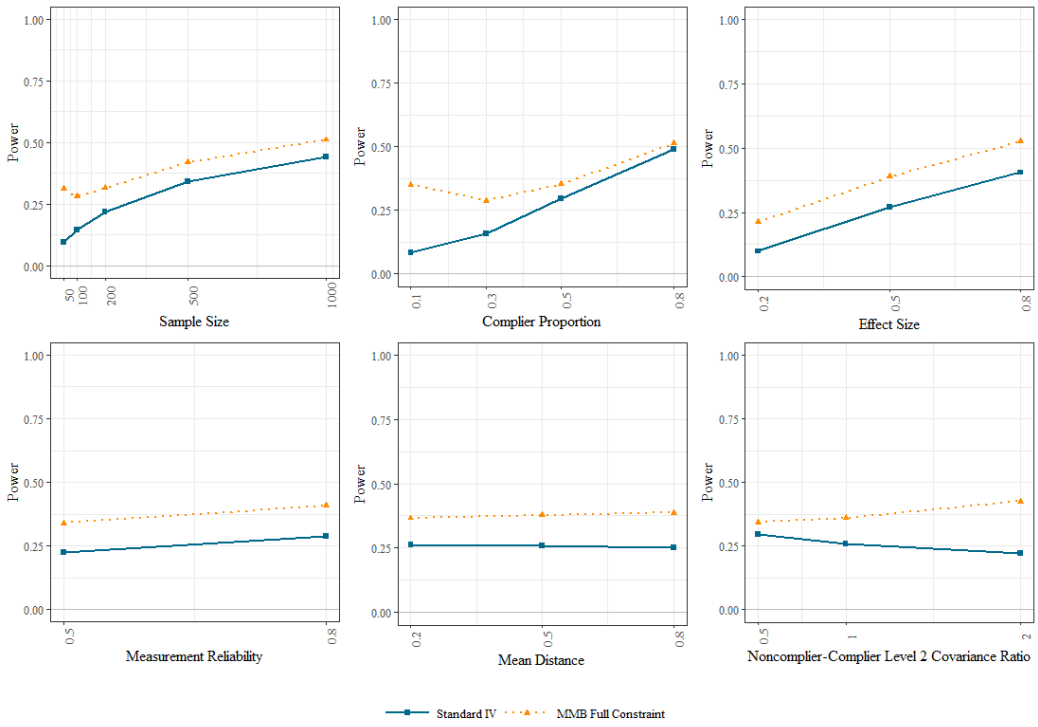

Figure 24. Mean values of power at each level of the factors for the SIV and the MMB-FC approaches

Table 22 gives a summary of the mean values of power over different levels of the six factors, and Figure 24 provides a visual presentation.

Overall, the SIV approach had a mean value of 0.2576. The three factors identified in the factorial ANOVA analysis, *PC*, *n* and *d*, displayed evident variations across their sublevels. All three factors showed a positive effect on the empirical power: higher proportion of compliers, higher sample size, or higher effect size gave rise to higher power. When *PC* increased from 0.1 to 0.8, the mean values of the power changed from 0.0832 to 0.4911 with a 0.4097 increase. While for the Sample Size factor, the mean values rose from 0.0946 to 0.4420, displaying a 0.3474 growth.

When d = 0.2, the mean value was only 0.0988. With d increased to 0.8, the mean value reached 0.4052, showing a 0.3064 increase.

The other three factors also manifested some variations across their sublevels, but their impacts were much milder. Measurement Reliability had a positive impact as well. The *var* factor and the *md* factor both had small negative effects on the empirical power.

The MMB-FC approach had higher power on average, 0.3760. When only looking at the main effects, the Effect Size factor manifested the greatest impact on the power, and the impact was positive. Increasing d from 0.2 to 0.8 caused the average values of power to rise from 0.2124 to 0.5261 (a 0.3137 increase). The positive trend was also true for Measurement Reliability, Mean Distance, and Noncomplier-Complier Level 2 Covariance Ratio, but their influences were on a much smaller scale.

Factors n and PC both had a more complex influence on power. For both factors, the impact was negative at first and then turned to positive. When having a really low sample size of 50, the mean value was 0.3123, but decreased to 0.2812 if increasing n to 100. If n kept increasing, the mean value stopped dropping and started to climb until reaching 0.5106 when n = 1,000. The difference between the largest and smallest mean values was 0.2294. Analogous pattern was found for the PC factor. With PC = 0.1, the mean value of power was 0.3505, but when PC = 0.3, the mean value reached the lowest point, 0.2870. When PC = 0.8, the value hiked up to 0.5132. The difference between the two mean values when PC = 0.3 and 0.8 was 0.2262.

As the factorial ANOVA analysis suggested, there was a "medium" sized interaction between the Sample Size factor and the Complier Proportion factor. The following section explores more on this interaction effect.

Table 23

Mean Values of Power by Different Configurations of PC and N for the SIV and the MMB_FC Approaches

n	PC	SIV	MMB-FC
50	0.1	**0.1050**	**0.4902**
	0.3	**0.0461**	**0.3230**
	0.5	**0.0770**	**0.2177**
	0.8	**0.1614**	**0.2040**
100	0.1	**0.0955**	**0.4241**
	0.3	**0.0562**	**0.2179**
	0.5	**0.1338**	**0.1897**
	0.8	**0.2897**	**0.2959**
200	0.1	**0.0782**	**0.3564**
	0.3	**0.1142**	**0.2005**
	0.5	**0.2237**	**0.2502**
	0.8	**0.4499**	**0.4558**
500	0.1	**0.0680**	**0.2797**
	0.3	**0.2143**	**0.2839**
	0.5	**0.4157**	**0.4396**
	0.8	**0.6622**	**0.6708**
1000	0.1	**0.0764**	**0.2588**
	0.3	**0.3366**	**0.4040**
	0.5	**0.5779**	**0.5953**
	0.8	**0.7766**	**0.7829**

Note. In the same row, the number with a bigger absolute value was bold, and a smaller absolute value was bold and underlined.

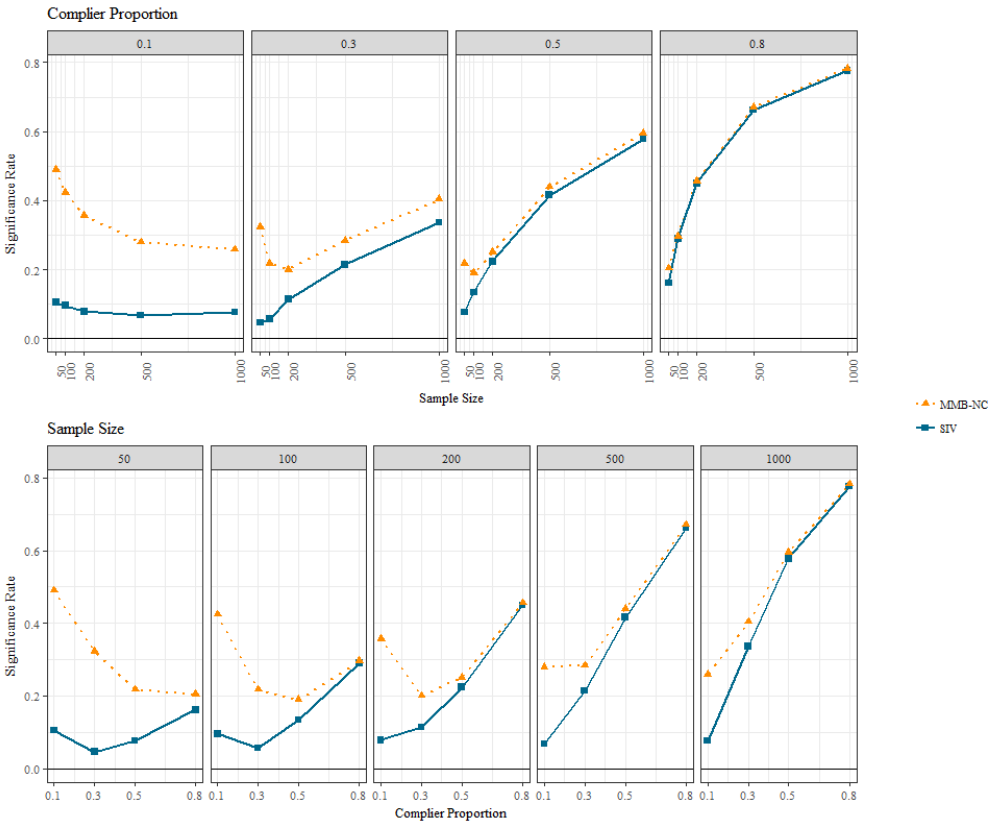

Figure 25. Mean values of power by different configurations of *PC* and *n* for the SIV and the MMB_FC approaches.

Table 23 summarizes the mean values of power by different configurations of *PC* and *n*. Figure 25 displays the interaction effect with two rows. For both estimation approaches, their lines changed noticeably across different plots within a single row. The change was more salient for the MMB-FC approach, but it was definitely very evident for the SIV approach, too. Checking the factorial ANOVA result, the interaction term of *n*PC* was found to have a pseudo partial η^2 of 4.35% for the SIV approach. Although it did not meet the 6% cutoff criterion, it was big enough to demonstrate itself in Figure 25.

For the SIV approach, the main effects of both *n* and *PC* were positive. However, after taking the interaction effect into consideration, the effects of both factors changed across different levels of the controlling factor. Specifically, with *PC* = 0.1, Sample Size had a negative effect first until it changed from 500 to 1,000. In addition, the variation among the five *n* levels within this *PC* level was very small: 0.0680 to 0.1050 (0.0370 difference). With higher *PC* values, Sample Size demonstrated a clear positive effect, and the effect was the strongest with *PC* = 0.8, increasing from 0.1614 to 0.7766 with a change of 0.6152.

The trend of the *PC* factor also showed considerable fluctuation across different levels of *n*. When *n* = 50 and 100, *PC* first had a negative effect on the power when changing 0.1 to 0.3, but the effect became positive when *PC* continued to increase from 0.3. With *n* = 50, the difference between the highest (0.1614) and lowest (0.0461) mean values was the smallest, 0.1153. With higher *n* levels, *PC* always demonstrated a positive influence and the influence became stronger with higher *n*. With *n* = 1,000, the mean values increased by 0.7002 from 0.0764 to 0.7766.

The MMB-FC approach had very comparable patterns as the SIV approach, only with more noticeable pattern changes across sublevels. In detail, with *PC* = 0.1, *n* had a negative effect, and the variation among the five sample size levels within this proportion level was quite sizeable, decreasing from 0.4902 to 0.2588 with a 0.2314 decrease. With *PC* = 0.3 and 0.5, the effect of *n* became curvature. The mean values decreased first as *n* became larger, but this trend was turned around to become

positive with continuous increase of *n*. The only difference was that the lowest mean value occurred when *n* = 200 with *PC* = 0.3 and when *n* = 100 with *PC* = 0.5. With *PC* = 0.8, the effect of *n* was purely positive: the mean values changed from 0.2040 to 0.7829 (0.5789 increase).

The trend change of the *PC* factor was similar to that of the *n* factor. When *n* = 50, *PC* had a negative effect. With *n* = 100, 200, and 500, the effect of *PC* was curvature: decreasing first and then increasing. The only difference was the turning point. With *n* = 1,000, the influence of *PC* became completely positive. The effect of *PC* culminated with *n* = 1,000, creating a difference of 0.5241 by changing the mean values of power from 0.2588 to 0.7829.

Inspecting the two estimation methods together, the MMB-FC method on average yielded a much higher power, 0.3760, than the SIV approach, 0.2576. Across the sublevels within each factor, the MMB-FC approach always had a higher mean. This was clear in Table 22, where the numbers under the "MMB Full Constraint" header were all bold, indicating being the larger values of the two estimation methods. This was also true for the interaction table too. Within each configuration of *PC* and *n*, the MMB-FC approach always yielded higher mean values of power than the SIV approach. Consequently, in Table 23, the numbers under the "MMB Full Constraint" header were all bold. The difference between the two methods diminished to almost negligible with *PC* = 0.8.

Results of factor effects with applicable conditions. This section examined the empirical power for the MMB-NC approach using applicable simulation conditions.

Table 24

Factorial ANOVA Result: Power with the MMB-NC Method

		Max Pseudo Partial Eta Squares			
Estimation Approach	**Term**	**Sample 1**	**Sample 2**	**Sample 3**	***Mean of All Maxes***
MMB No Constraint	Effect Size	**30.70**	**31.07**	**32.53**	***31.44***
	Complier Proportion	**15.70**	**16.51**	**16.95**	***16.38***
	Sample Size	4.77	4.68	4.90	*4.78*

Note. $\eta^2 \geq 14\%$: bold and underlined; $6\% \leq \eta^2 < 14\%$: bold; $\eta^2 < 6\%$: grayed out.

Table 24 organizes the results of the factorial ANOVA analyses for the MMB-NC estimation method with sample units under applicable conditions. Two terms met the 6% criterion and were hence selected. The "Effect Size" term was the most influential one and had a "Large" effect, accounting for 30.70% of the sum of the deviance reduced by this term plus the residual deviance. The "Complier Proportion" term had the second largest pseudo partial η^2s, 15.70%.

Table 25

Mean Values of Power at each Level of the Factors for the Three Estimation Approaches under Applicable Conditions

		Standard IV	MMB No Constraint	MMB Full Constraint
Overall	Overall	**0.4972**	**0.5441**	0.5298
Sample Size	500	**0.4306**	0.4625	**0.4651**
	1000	**0.5637**	**0.6094**	0.5941
Complier Proportion	0.3	**0.2755**	**0.3539**	0.3442
	0.5	**0.4968**	**0.5664**	0.5176
	0.8	**0.7195**	**0.7498**	0.7270
Effect Size	0.2	**0.1762**	0.1925	**0.2248**
	0.5	**0.5517**	**0.6101**	0.5907
	0.8	**0.7638**	**0.8300**	0.7738
Measurement Reliability	0.5	**0.4355**	0.4617	**0.4716**
	0.8	**0.5588**	**0.5980**	0.5876
Mean Distance	0.2	0.5194	**0.5530**	**0.5135**
	0.5	**0.5011**	**0.5442**	0.5290
	0.8	**0.4713**	0.5355	**0.5470**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	0.5572	**0.5948**	**0.4058**
	1	**0.5026**	0.5041	**0.5259**
	2	**0.4319**	0.5314	**0.6572**

Note. In the same row, the number with the biggest absolute value was bold, and the smallest absolute value was bold and underlined.

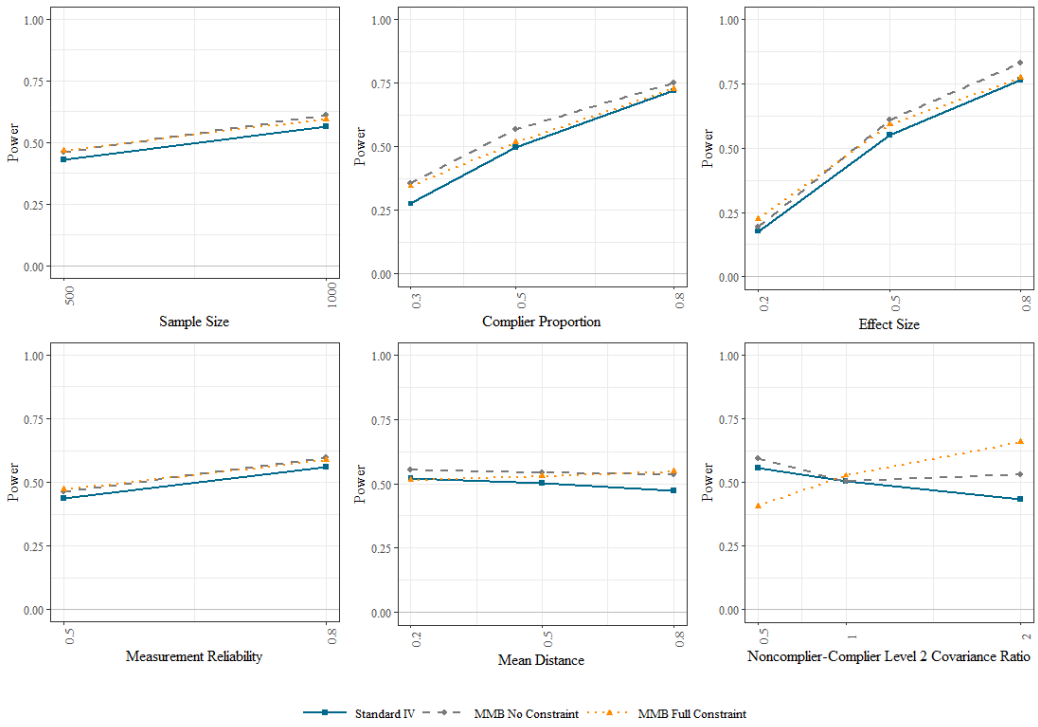

Figure 26. Mean values of power at each level of the factors for the three estimation approaches under Applicable Conditions.

Table 25 sums up the mean values of power over different levels of the six factors for the MMB-NC method, and Figure 26 is a graphical representation of Table 25. For comparison, Table 25 and Figure 26 also include the other two estimation methods' results using only sample units under applicable conditions.

Overall, 54.41% of all used datasets yielded a significant result. The two terms pinpointed by the factorial ANOVA analyses demonstrated obvious positive effects: higher effect size or higher complier proportion led to higher mean. When *d* changed from 0.2 to 0.8, the mean values increased from 0.1925 to 0.8300, exhibiting an increase of 0.6375. When *PC* rose from 0.3 to 0.8, the mean climbed from 0.3539 to 0.7498 with a 0.3959 growth.

The Sample Size factor and the Measurement Reliability factor also manifested positive effect on the power with a much smaller scale. Mean Distance had a very small negative effect. For the *var* factor, the lowest rate occurred with *var* = 1, and the highest with *var = 0.5*. The effect was also negligible.

In summary, under the Applicable Conditions, the MMB-NC method had the highest average power, 0.5441, the MMB-FC approach had a slightly smaller mean value, 0.5298, and the SIV method had the smallest mean value, 0.4972. The difference among the three overall mean values was not too evident. For most sublevels, either the MMB-NC or the MMB-FC method yielded the highest average power, and the SIV method yielded the lowest. As shown in Table 25, most numbers under the "Standard IV" header are bold and underlined, indicating having the lowest absolute values among the three numbers in a row. Most numbers under either the "MMB-NC" header or the "MMB-FC" header are just bold, indicating having the highest absolute values.

Guidance on choosing n and rel with respect to power.

Table 26

Percentages of Cells with Cell Average Empirical Power Meeting the Satisfactory Criterion[a]

		Standard IV												MMB No Constraint										MMB Full Constraint											
		PC=0.1			PC=0.3			PC=0.5			PC=0.8			PC=0.1			PC=0.3			PC=0.5			PC=0.8			PC=0.1			PC=0.3			PC=0.5			PC=0.8		
n	rel	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8	d=0.2	d=0.5	d=0.8
50	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	88.89	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	***100.00***
500	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	***100.00***	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	11.11	0.00	0.00	***100.00***	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.22	0.00	0.00	***100.00***
500	0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.67	0.00	***100.00***	***100.00***	NA	NA	NA	0.00	0.00	0.00	0.00	0.00	***100.00***	0.00	***100.00***	***100.00***	0.00	0.00	0.00	0.00	0.00	22.22	0.00	11.11	66.67	0.00	***100.00***	***100.00***
1000	0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	77.78	0.00	***100.00***	***100.00***	NA	NA	NA	0.00	0.00	11.11	0.00	0.00	***100.00***	0.00	***100.00***	***100.00***	0.00	0.00	0.00	0.00	11.11	22.22	0.00	22.22	77.78	0.00	***100.00***	***100.00***
1000	0.8	0.00	0.00	0.00	0.00	0.00	33.33	0.00	44.44	***100.00***	0.00	***100.00***	***100.00***	NA	NA	NA	0.00	0.00	66.67	0.00	88.89	***100.00***	0.00	***100.00***	***100.00***	0.00	0.00	0.00	0.00	22.22	55.56	0.00	66.67	88.89	0.00	***100.00***	***100.00***

Note. Percentages equal to 100% were italicized, bold, and underlined. *NA* were conditions excluded for the MMB-NC method. All designs with $d = 0$ were excluded.

[a]"Satisfactory Criterion" meant that the cell average empirical power was bigger than or equal to 80%.

To provide guidance on choosing acceptable sample size and measurement reliability combinations to yield adequate power, Table 26 uses cross tabulation to summarize the significant rates with respect to the three factors identified above (i.e., *n*, *PC*, and *d*). Only designs with $d > 0$ are used in this table, so this table is solely about empirical power.

Each number in this table is a summary of the nine cells with the same configuration of *n*, *PC*, *d*, and *rel*. Each number represents the percentage of cells with cell average empirical power higher than or equal to the pre-specified satisfactory power rate, 80%. If all cells within a configuration have empirical power bigger than or equal to 80%, the number representing this configuration in Table 26 will be 100% and is italicized, bold, and underlined. Note that for the MMB-NC method, configurations excluded from the analyses are labeled with NA.

For the SIV approach, when *PC* was 0.1 or 0.3, none of the design cells could yield a satisfactory empirical power, since the numbers in these columns are all 0. Even with $PC = 0.5$, only the most optimistic combination of *n*, *rel*, and *d* could guarantee that all cells nested within would reach 80%. However, when *PC* reaches 0.8, $n = 500$ and $rel = 0.5$ could guarantee all cells having empirical power bigger than 80%, if $d = 0.8$. With lower effect size of 0.5, the configurations should be either $n = 500$ and $rel = 0.8$ or just $n = 1{,}000$.

The MMB-NC approach had similar results as the SIV approach when $PC = 0.8$. However, this estimation method was more lenient with the situation when

complier proportion was 0.5. With the biggest effect size and the more reliable measurement, a sample size of 500 was enough for guaranteeing satisfactory empirical power for all cells nested within. A sample size of 1,000 was required if the measurement reliability was 0.5.

The MMB-FC approach had similar results as the SIV approach when complier proportion was 0.8. The only difference was that the configuration of n = 200, rel = 0.8, d = 0.8, and PC = 0.8 could also guarantee empirical power for the cells nested within. However, when PC = 0.5, none of the available configurations would be acceptable.

4.4.2. Type I error

Results of factor effects with all conditions. This section examined the empirical type I error rate for the SIV and the MMB-FC approaches using all simulation conditions.

Table 27

Factorial ANOVA Result: Type I Error with the SIV and the MMB-FC Methods

		Max Pseudo Partial Eta Squares			
Estimation Approach	**Term**	**Sample 1**	**Sample 2**	**Sample 3**	***Mean of All Maxes***
Standard IV	Sample Size*Complier Proportion	1.65	1.87	1.67	*1.73*
	Complier Proportion	1.38	1.28	1.38	*1.35*
	Mean Distance*Sample Size*Complier Proportion	0.38	0.57	0.46	*0.47*
MMB Full Constraint	Complier Proportion	**9.08**	**9.99**	**8.53**	***9.20***
	Sample Size	1.94	2.06	1.84	*1.94*
	Noncomplier-Complier Level 2 Covariance Ratio	1.29	1.26	1.29	*1.28*

Note. $6\% \leq \eta^2 < 14\%$: bold; $\eta^2 < 6\%$: grayed out.

Table 27 organizes the results of the factorial ANOVA analyses for the SIV and the MMB-FC estimation methods with sample unites under all conditions. In terms of type I error rate, none of the terms was important enough for the SIV approach. For the MMB-FC approach, the Complier Proportion showed "Medium" effect, accounting for 9.20% of the sum of the deviance reduced by this term plus the residual deviance. The following discussion only explores the main effects of the six factors.

Table 28

Mean Values of Type I Error at each Level of the Factors for the SIV and the MMB-FC Approaches

		Standard IV	MMB Full Constraint
Overall	Overall	0.0441	**0.1518**
Sample Size	50	0.0473	**0.2461**
	100	0.0467	**0.1695**
	200	0.0429	**0.1338**
	500	0.0425	**0.1184**
	1000	0.0418	**0.1254**
Complier Proportion	0.1	0.0653	**0.3084**
	0.3	0.0304	**0.1600**
	0.5	0.0345	**0.0848**
	0.8	0.0471	**0.0551**
Measurement Reliability	0.5	0.0435	**0.1416**
	0.8	0.0445	**0.1603**
Mean Distance	0.2	0.0415	**0.1354**
	0.5	0.0434	**0.1509**
	0.8	0.0473	**0.1690**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	0.0464	**0.1950**
	1	0.0430	**0.1185**
	2	0.0428	**0.1432**

Note. In the same row, the number with the biggest absolute value was bold, and the smallest absolute value was bold and underlined.

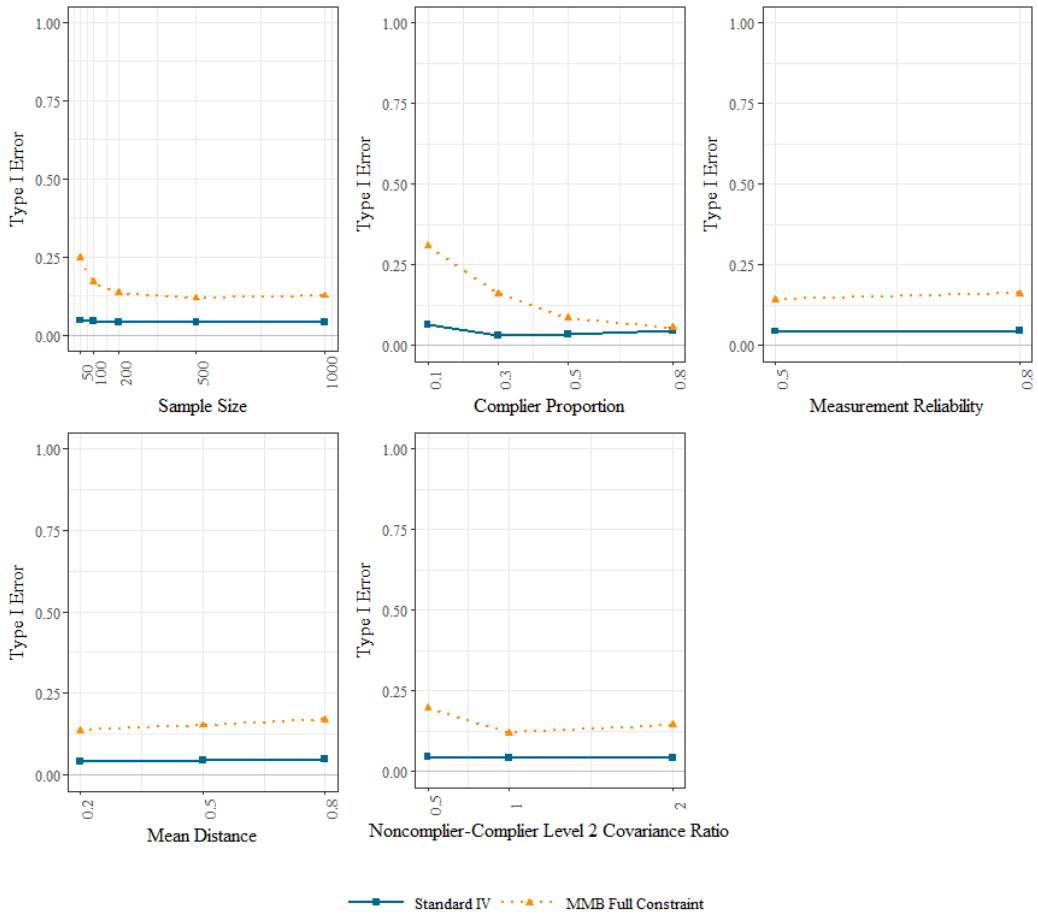

Figure 27. Mean values of type I error at each level of the factors for the SIV and the MMB-FC approaches

Table 28 gives a summary of the mean values of type I error over different levels of the six factors, and Figure 27 provides a visual presentation.

Overall, the SIV approach had a mean value of 0.0441. Similar to the ANOVA analysis results, none of these factors displayed evident effects on the type I error rate. The MMB-FC approach had a much higher mean value of type I error rate, 0.1518. Among the six main effects, the Complier Proportion factor had the greatest impact on the type I error rate. Increasing *PC* from 0.1 to 0.8 caused the average type I error rate to decrease from 0.3084 to 0.0551 (a 0.2533 decrease). The same trend was also

true for the Sample Size factor, but its influence was on a much smaller scale. The effects of the other four factors were negligible.

Inspecting the two estimation methods together, the SIV approach on average had closer to 0.05 type I error rate while the MMB-FC method on average yielded a much higher type I error rate, 0.1518. Across the sublevels within each factor, the MMB-FC approach always had a higher mean value too.

Results of factor effects with applicable conditions. This section examined the type I error rate for the MMB-NC method using applicable simulation conditions.

Table 29

Factorial ANOVA Result: Type I Error with the MMB-NC Method

		Max Pseudo Partial Eta Squares			
Estimation Approach	**Term**	**Sample 1**	**Sample 2**	**Sample 3**	***Mean of All Maxes***
MMB No Constraint	Mean Distance*Sample Size*Complier Proportion	0.33	NA	NA	*0.33*
	Complier Proportion*Noncomplier-Complier Level 2 Covariance Ratio	0.29	0.46	0.18	*0.31*
	Mean Distance*Sample Size*Noncomplier-Complier Level 2 Covariance Ratio	0.29	NA	NA	*0.29*

Note. $\eta^2 < 6\%$: grayed out.

29 organizes the results of the factorial ANOVA analyses for the MMB-NC estimation method using applicable conditions. None of the terms was important enough for this approach. The following discussion only explores the main effects of the six factors.

Table 30

Mean Values of Type I Error at each Level of the Factors for the MMB-NC Approach under Applicable Conditions

		Standard IV	MMB No Constraint	MMB Full Constraint
Overall	Overall	**0.0450**	0.0486	**0.0891**
Sample Size	500	**0.0428**	0.0504	**0.0780**
	1000	**0.0471**	**0.0471**	**0.1000**
Complier Proportion	0.3	**0.0401**	0.0474	**0.1415**
	0.5	**0.0452**	0.0481	**0.0755**
	0.8	**0.0496**	**0.0506**	0.0504
Measurement Reliability	0.5	**0.0460**	0.0464	**0.0858**
	0.8	**0.0440**	0.0500	**0.0923**
Mean Distance	0.2	**0.0456**	0.0468	**0.0552**
	0.5	**0.0446**	0.0487	**0.0852**
	0.8	**0.0447**	0.0502	**0.1268**
Noncomplier-Complier Level 2 Covariance Ratio	0.5	**0.0453**	0.0500	**0.1166**
	1	**0.0432**	0.0477	**0.0491**
	2	**0.0464**	0.0480	**0.1016**

Note. In the same row, the number with the biggest absolute value was bold, and the smallest absolute value was bold and underlined.

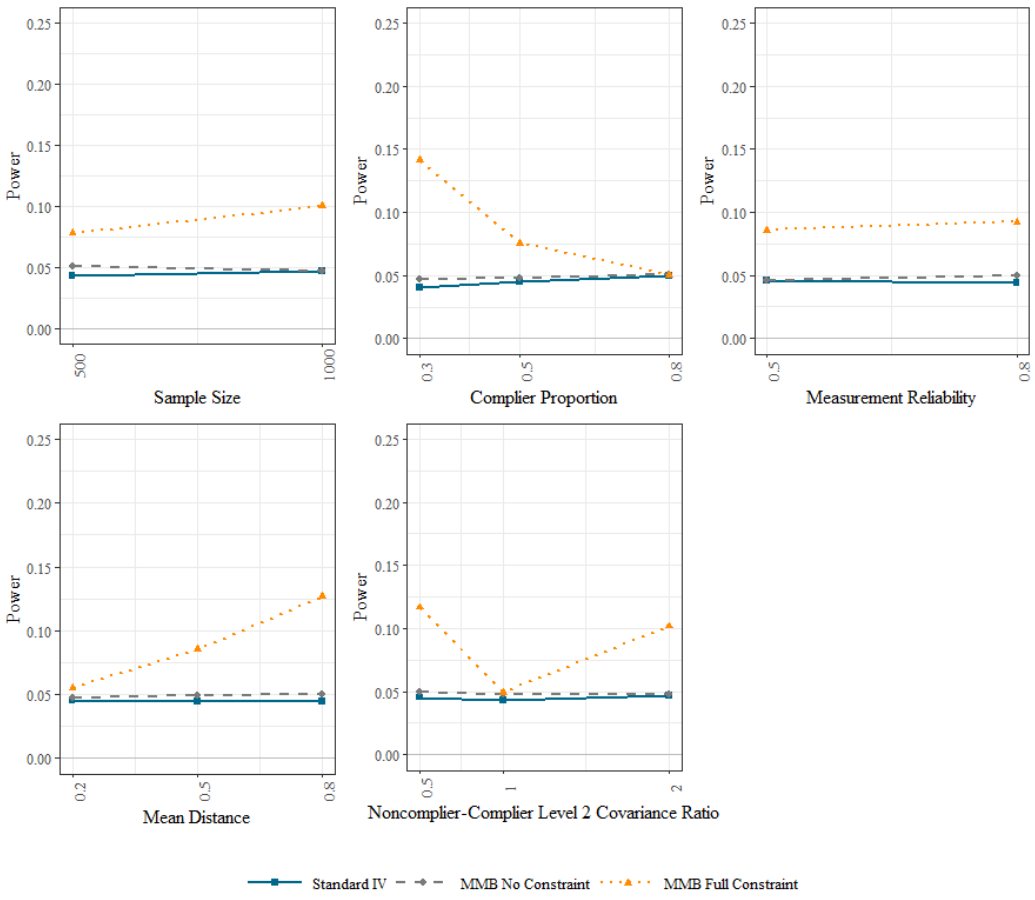

Figure 28. Mean values of type I error at each level of the factors for the MMB-NC approach under Applicable Conditions

Table 30 gives a summary of the mean values of type I error over different levels of the six factors, and Figure 28 provides a visual presentation. As suggested by the ANOVA results, none of these factors demonstrated a clear impact on the type I error rate of the MMB-NC approach.

Comparing the three estimation methods together under the applicable conditions, the MMB-NC method on average yielded a type I error rate (0.0486) that was the closest to the pre-specified significance level of 0.05. The SIV method had slightly more under estimation, resulting in an average type I error rate of 0.0450. The

MMB-FC method had a much higher type I error rate, 0.0891. Across the sublevels within each factor, the same conclusion remained.

Guidance on choosing n and rel with respect to type I error rate.

Table 31

Percentages of Cells with Cell Average Empirical Type I Error Rate Meeting the Negligible Criterion[a]

		Standard IV				MMB No Constraint				MMB Full Constraint			
n	**rel**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**	**PC=0.1**	**PC=0.3**	**PC=0.5**	**PC=0.8**
50	0.5	0.00	0.00	0.00	11.11	NA	NA	NA	NA	0.00	0.00	0.00	11.11
50	0.8	0.00	11.11	0.00	33.33	NA	NA	NA	NA	0.00	0.00	0.00	0.00
100	0.5	0.00	0.00	0.00	44.44	NA	NA	NA	NA	0.00	0.00	22.22	55.56
100	0.8	0.00	0.00	0.00	22.22	NA	NA	NA	NA	0.00	0.00	0.00	11.11
200	0.5	11.11	0.00	0.00	55.56	NA	NA	NA	NA	0.00	11.11	22.22	66.67
200	0.8	0.00	0.00	11.11	44.44	NA	NA	NA	NA	0.00	0.00	33.33	11.11
500	0.5	33.33	33.33	44.44	77.78	NA	33.33	33.33	44.44	0.00	11.11	33.33	66.67
500	0.8	0.00	0.00	33.33	22.22	NA	22.22	55.56	33.33	0.00	22.22	22.22	44.44
1000	0.5	0.00	33.33	44.44	44.44	NA	22.22	33.33	33.33	0.00	11.11	22.22	33.33
1000	0.8	11.11	55.56	55.56	33.33	NA	33.33	44.44	66.67	0.00	33.33	33.33	66.67

Note. Percentages equal to 100% were italicized, bold, and underlined. NA were conditions excluded for the MMB-NC method. All designs with $d = 0$ were excluded.

[a]"Negligible Criterion" meant that the cell average empirical type I error rate was within the bound of 4.5% and 5.5%.

Table 31**Error! Reference source not found.** summarizes the empirical type I error rate with respect to Sample Size, Complier Proportion, and Measurement Reliability. Only designs with $d = 0$ are used for this table.

Each number represents the percentage of design cells with cell average empirical type I error rate within the negligible bounds, 4.5% and 5.5%, among the

nine cells with the same configuration of the *n*, *rel*, and *PC*. None of the numbers in Table 31 was 100, indicating that none of these configurations could guarantee that all cells nested within had an average type I error rate within the 4.5−5.5% bounds.

This chapter presents detailed analysis result with respect to the five research questions. The next chapter will summarize the result, highlight key findings, make connections among the findings of the five research questions, and prove implications for future studies.

Chapter 5: Discussion

Despite the popularity of using the LGMs for longitudinal experiments and the wide recognition of the noncompliance issue accompanying random experiments, addressing noncompliance while using the LGMs for longitudinal experiments is a relatively new approach. Although previous studies (Jo & Muthén, 2001, 2003) have explored how to use the mixture model based approach in such a scenario, the functionality of this approach needs more investigation, especially with three compliance classes. In addition, the SIV method is also commonly utilized for studies with noncompliance problems. It can be readily applied to the LGMs for the CACE estimation. The estimation of the treatment effect is asymptotically unbiased, and the standard errors can be empirically estimated with the bootstrap technique. The two approaches are asymptotically equivalent, but their performance may differ under different conditions. For example, one method can be more robust to extreme conditions than the other.

The aim of this dissertation is to examine the functionality of the MMB and the SIV approaches in estimating the longitudinal CACE within the LGM framework with respect to a wide range of factors. The LGM chosen had four measurement points exhibiting a linear growth. There were three compliance classes representing compliers, always-takers, and never-takers. The effects of six factors on the treatment effect estimation using either the SIV or the MMB approach were examined. The MMB approach had two variants: the MMB-NC method estimated

α_{Int}, α_{Slp}, γ, σ_{ζ}^2, and σ_{ε}^2 for each compliance class separately, and the MMB-FC method only allowed the α_{Int}, α_{Slp}, and γ to be freely estimated and constrained σ_{ζ}^2 and σ_{ε}^2 to be the same across different compliance classes. The estimation quality was evaluated with four main criteria: 1) estimation success rate, 2) estimation accuracy, 3) empirical power, and 4) empirical type I error rate.

The six simulation factors were Sample Size, Complier Proportion, Effect Size, Measurement Reliability, Mean Distance, and Noncomplier-Complier Level 2 Covariance Ratio. There were six levels for the Sample Size factor, 50, 100, 200, 500, and 1,000, four levels for the Complier Proportion factor, 0.1, 0.3, 0.5, and 0.8, four levels for the Effect Size factor, 0, 0.2, 0.5, and 0.8, two levels for the Measurement Reliability factor, 0.5 and 0.8, three levels for the Mean Distance factor, 0.2, 0.5, and 0.8, and three levels for the Noncomplier-Complier Level 2 Covariance Ratio factor, 0.5, 1, and 2. All six factors were fully crossed resulting in 1,440 configurations in total. Within each configuration, 1,000 replications were used. Low complier proportion was specifically included in this study because it is fairly common for big intervention studies to have low compliance rate. For example, the Angrist and Krueger's (1991) study of the causal effect of education on earnings was found to have compliance rate of only 2%.

The estimation success rate was investigated first because non-successful estimations were likely to be more prevalent for the MMB-NC estimation approach than the other two. After examining the successful estimation rate, the author found that for most simulation conditions, the MMB-NC approach had an extremely low

success rate, so for the other dependent variables, this approach was only examined under simulation conditions that led to all design cells having at least 300 successful estimations. These conditions were called the "Applicable Conditions".

The estimation accuracy was examined next with two bias measures, the parameter estimation bias and the standard error estimation bias. Each measure was analyzed using both Simple Bias and Relative Bias.

In the end, the statistical power and type I error rate were evaluated together by using the Significant Indicator variable because both criteria could be described as significant rate. When the true effect size was not zero, the significant rate was the empirical statistical power; when the true effect size was zero, the significant rate was the empirical type I error rate.

5.1.Summary of Factor Effects

This section summarizes the findings. Note that for the MMB-NC approach, only applicable conditions were used.

5.1.1. Findings on success rate

For all three estimation approaches, Sample Size and Measurement Reliability were two dominant factors for successful estimations. With higher *n* and/or higher *rel*, the probability of obtaining a successful estimation was higher. This result is consistent with the findings from previous studies (Gagne & Hancock, 2006; Tolvanen, 2007). In addition, for the MMB-NC approach, Complier Proportion was the second most important factor for estimation success, more influential than

Measurement Reliability. *PC* = 0.5 yielded higher average success rate than other *PC* levels. All influential terms for the three estimation approaches were main effect terms, indicating that their effects do not vary too much across different levels of other factors.

Comparing the three approaches, the SIV method yielded the highest overall mean success rate and the highest individual level mean success rates within each factor. The MMB-FC method had slightly lower means, but the difference between the two estimation methods became minimum with higher *n* or higher *rel*. The MMB-NC method, on the other hand, resulted in a much lower overall mean success rate and individual level mean success rates within each factor. The gap between this estimation method and the other two did not diminish to a negligible level by changing only one simulation factor. In fact, only with very favorable conditions on *n*, *PC*, and *rel* (i.e., n = 1,000, $PC \geq 0.3$, and *rel* = 0.8 or $n \geq 500$, $0.3 \leq PC \leq 0.5$, and *rel* = 0.8) did the MMB-NC method yield an average success rate close to the other two.

This finding is consistent with the study by Tolvanen (2007), where more complex models were followed with lower success rates. In the study of Tolvanen (2007), a Mean Distance equivalent factor was found to have a noticeable positive effect on estimation success rate, but the current study finds that the change of the mean success rates across different levels of *md* is minimum. The main reason is that the 2007 study included a much wider range of *md* values, ranging from 0.5 to 5, whereas the current study limits the *md* values to more realistic choices, 0.2 to 0.8. With the current choices, a small positive effect of *md* was observed for the two

MMB estimation approaches, but the increase was too small to be picked up by the factorial ANOVA analysis.

5.1.2. Findings on estimation bias

Parameter estimation bias. For all three estimation methods, none of the terms used in the factorial ANOVA analyses was kept, suggesting that there is too much variation on parameter estimation and the ANOVA model can barely explain anything. This result is consistent with Tolvanen's 2007 study, where the proportion of the squared bias of the method was found to be extremely small.

For the MMB-FC approach, *PC*var* was the most influential term for both estimation bias measures, and it was the only term that met the 2%, "small" effect, criterion, although it was actually smaller than the cutoff criterion, 6%. The interaction plot shows that smaller *var* value led to smaller estimated treatment effect: when *var* = 0.5, there was on average underestimation; when *var* = 1 or 2, there was on average overestimation, and *var* = 2 yielded a larger overestimation overall. The difference among the three *var* levels shrinked with higher *PC*. In conclusion, for the MMB-FC method, the model was misspecified when *var* = 0.5 or 2, and the misspecification can severely distort the estimation. The distortion is more exacerbated when combining with low *PC*.

Imbens and Rubin (1997b) used a very small simulation study to generate cross-sectional data and found that the MMB method yielded a very small negative bias while the SIV approach yielded a much bigger positive bias. The current study

used a different model and manipulated a much wider range of factors. As a result, different results are found.

Comparing the SIV and the MMB-FC approaches using all conditions, the MMB-FC method on average yielded a positive and lower magnitude bias, while the SIV approach overall produced a negative bias with a slightly bigger magnitude. However, the MMB-FC method was much more susceptible to unfavorable choices of n and PC (i.e., $n < 200$ and $PC = 0.1$). When $n \geq 200$ or $PC > 0.1$, the MMB-FC method yielded sublevel mean values of bias with slightly smaller magnitudes than the SIV approach. In addition, the MMB-FC method had sublevel mean values of bias with higher magnitudes, when $var = 0.5$ or 2, while the SIV method was not much affected by the change of *var*. The two approaches did not differ much on the magnitudes of sublevel mean values of bias with respect to *d*, *rel*, and *md*.

While comparing the three estimation methods under applicable conditions, all three methods on average yielded a small-sized negative bias. The MMB-NC method was the least biased, the SIV approach was slightly more biased, and the MMB-FC method was the most biased. For most sublevels within each factor, the same conclusion still held. Their differences diminished to a minimum when $PC = 0.8$.

Standard error estimation bias. In terms of the Simple SE Bias, *PC* and *n* were the two important factors for the SIV and the MMB-FC methods. The main effect of *n* and the interaction effect of *n***PC* were both selected by the two approaches. Because the two terms had very similar effects on both estimation approaches, the following description applies to both of them.

The main effect of *n* on Simple SE Bias was positive for both approaches. As the average Simple SE Bias values across the sub-levels of *n* were all negative, the positive effect means that higher *n* value leads to closer-to-zero mean values. After taking *PC* into account, the effect of *n* changed considerably across different *PC* levels for both estimation approaches. Low *n* and low *PC* together yielded large negative mean values. With either high *n* or high *PC*, the bias became much less prominent.

While with regard to Relative SE Bias, the factorial ANOVA analyses selected one term for the SIV approach, the interaction effect of *n***PC*, and selected none for the MMB-FC method. The conclusions on the interaction effect were the same for the Relative SE Bias and the Simple SE Bias.

None of the terms was influential enough for the MMB-NC method. The reason that the MMB-NC approach was not greatly affected by the simulation factors is probably because that the analyses only included two big *n* levels, 500 and 1,000. With such high *n* values, none of the other factors can make too much change.

Comparing the SIV and the MMB-FC approaches using all conditions, both methods on average underestimate the SE, but the MMB-FC approach yielded more

underestimation. This was true using both Simple and Relative SE Bias. Within each sublevel of a factor and within each combination of different *n* and *PC* levels, the MMB-FC approach also yielded more underestimation than the SIV approach, except when *PC* = 0.1 and *n* = 50 or 100.

While comparing the three estimation methods under applicable conditions, all three methods on average yielded a positive bias with a small magnitude. The MMB-NC method had a slightly larger bias than the other two, and the MMB-FC method had a slightly closer to zero overall bias. For most sublevels, the same order persevered. The only exception was when *var* = 0.5, the SIV approach had the biggest mean value of bias. The difference among the three estimation methods was extremely small and also diminished with higher sample size, higher complier proportion, and higher measurement reliability.

5.1.3. Findings on power and type I error rate

Empirical power and empirical type I error rate were both calculated as the significance rate. The only difference was that the former should use simulation designs with non-zero true effects while the latter should use designs with zero true effects.

For empirical power, *d* and *PC* were two important factors for all three estimation methods, and *n* was selected for the SIV and the MMB-FC approaches. The reason that the MMB-NC approach did not select *n* as an influential factor is probably because that the analysis for this estimation method only included two big *n*

levels, 500 and 1,000. As both *n* values were relatively high, there was not much variation.

For the SIV approach, the empirical power increased together with *PC*, *n* and *d*. The MMB-NC approach had a similar pattern regarding factor *d* and *PC*, but factor *n* did not show an effect as influential as *PC* and *d*. For the MMB-FC approach, its empirical power was also positively influenced by factor *d*, but the effect of *n* and *PC* was primarily shown through their interaction. With $PC = 0.1$, higher sample size unexpectedly led to lower power. When $PC = 0.3$ or 0.5, the power decreased at first and increased again along with the increase of *n*. When $PC = 0.8$, the power increased together with *n*. A similar pattern was observed when analyzing the effect of *PC* across different levels of factor *n*. When $n = 50$, higher complier proportion also led to lower power. With $n = 100$, 200 or 500, the trend decreased first and then increased. With $n = 1{,}000$, the power increased as *PC* became higher.

Inspecting the SIV and the MMB-FC approaches using all conditions, the MMB-FC method on average led to a higher power than the SIV approach. Within each sublevel of all factors and within each configuration of *n* and *PC*, the MMB-FC approach also yielded higher power. Their difference diminished to a minimum with $PC = 0.8$.

While only Applicable Conditions were used, the MMB-NC method, the MMB-FC approach, and the SIV method had the highest, the middle, and the lowest overall power respectively, but their differences were extremely small. For most

sublevels, either the MMB-NC or the MMB-FC method yielded the highest power, and the SIV method yielded the lowest.

For type I error rate, only *PC* demonstrated an obvious effect for the MMB-FC method. As *PC* became larger, the type I error rate became closer to the pre-specified value, 0.05. Comparing the three estimation methods, the MMB-FC approach always performed much worse than the other two methods, under all simulation conditions and under applicable conditions. This method yielded a much higher average type I error rate than the other two. Although the SIV method yielded an average type I error rate slightly further from 0.05 than the MMB-NC method under applicable conditions, the difference was almost negligible. Under all simulation conditions, the SIV method performed much better than the MMB-FC method.

Comparing the effects of these factors on power using the two MMB approaches to Jo's (2002) study, similar results are only found for the *var* and *md* factors. Note in order to be comparable with Jo's (2002) study, the results of the MMB-NC approach should be used for the *var* effect comparison and the results of the MMB-FC approach should be used for *md*. Although these factors are not selected by the factorial ANOVA analyses, the trends of the power across different levels of these two factors are the same as described in Jo's (2002) study. The power was the highest when $var = 0.5$ and lowest when $var = 1$. The power increased as *md* increased.

For other factors, Jo's (2002) findings do not hold in the current study. Jo (2002) found that using the MMB-FC model, *PC*, *n*, and *d* had a positive effect on the

power and the trends always remained positive even after taking the interaction effects into account. The current study reaches a similar conclusion for *d* (always positive), but very different conclusions for *PC* and *n* with respect to the MMB-FC approach. As described above, when inspecting the interaction effect of *n* and *PC*, if one factor had an extremely low value, higher value on the other factor unexpectedly led to lower power, if one factor had an medium size value, higher value on the other factor first led to lower power and then higher, and if one factor had relatively high value, the effect of the other factor became purely positive. The main reason for this difference is because the current study included two more *n* levels and one more *PC* level, and the added levels are all very unfavorable. If only using the conditions chosen by Jo (2002), the trends are similar.

5.1.4. Overall factor effects

Table 32 is a summary of all findings regarding factor effects, and Table 33 is a summary of the comparisons among the three approaches. Note that in Table 33 the MMB-NC approach is only comparable with the other two approaches under Applicable Conditions for all dependent variables. Therefore, there are two columns for approach comparisons, one for comparisons between the SIV and the MMB-FC approaches and one for comparisons among all three.

Table 32

Summary of Factor Effects

	Selected Terms	Term Effect
Success Indicator		
Standard IV	*n*	*n* ↑, success rate ↑
	rel	*rel* ↑, success rate ↑
MMB Full Constraint	*n*	*n* ↑, success rate ↑
	rel	*rel* ↑, success rate ↑
MMB No Constraint[a]	*n*	*n* ↑, success rate ↑
	PC	*n* ↑, success rate ↑ first then ↓. Peak at *PC*=0.5
	rel	*rel* ↑, success rate ↑
Simple Estimation Bias		
Standard IV	--	--

MMB Full Constraint	*PC***var*[b]	*var*=0.5: underestimation;
		var=1 or 2: overestimation, and *var*=2 has larger overestimation;
		low *PC* leads to more bias.

MMB No Constraint[a]	--	--
Relative Estimation Bias	--	
Standard IV	--	Same as Simple Estimation Bias
MMB Full Constraint	*PC***var*[b]	
MMB No Constraint[a]	--	
Simple SE Bias		
Standard IV	*n***PC*	Low *n* and low *PC* together, large negative bias means; High *n* or high *PC*, accurate estimation.
	n	*n* ↑, bias value ↑ and bias magnitude ↓
MMB Full Constraint	*n*	*n* ↑, bias value ↑ and bias magnitude ↓
	*n***PC*	Low *n* and low *PC* together, large negative bias means; High *n* or high *PC*, accurate estimation.
MMB No Constraint[a]	--	--

Table 32 (continued)

	Selected Terms	Term Effect
Relative SE Bias		
Standard IV	*n*PC*	Low *n* and low *PC* together, large negative bias means; High *n* or high *PC*, accurate estimation.
MMB Full Constraint	--	--
MMB No Constraint[a]	--	--
Power		
Standard IV	*PC*	PC↑, power↑
	n	n↑, power↑
	d	d↑, power↑
MMB Full Constraint	*n*PC*	*PC=0.1: n↑, power ↓;* *PC=0.3 or 0.5, n↑, power first ↓ then ↑;* *PC=0.8: n↑, power ↑.* *n=0: PC↑, power ↓;* *n=100, 200 or 500: PC↑, power first↓then ↑;* *n>200: PC↑, power ↑.*
	d	d↑, power↑
MMB No Constraint[a]	*d*	d↑, power↑
	PC	PC↑, power↑
Type I Error		
Standard IV	--	--
MMB Full Constraint	*PC*	*PC* ↑, type I error ↓
MMB No Constraint[a]	--	--

Note. -- Not exist or not applicable. [a]Only applicable conditions were used for this dependent variable.

[b]This term is smaller than the cutoff value of 6%, so it is not actually "selected". It is presented here because it is greater than 2%, and none of the other terms meet the 6% rule.

Table 33

Summary of Estimation Approaches Comparison

	Approach Comparison	
	All Conditoins	**Applicable Conditions**
Success Indicator		
Standard IV	Highest (overall and sublevels).	--
MMB Full Constraint	Slightly lower (overall and sublevels). Difference with Standard IV disappear with higher n and higher *rel*.	--
MMB No Constraint[a]	Much lower (overall and sublevels). Difference with the other two disappear with combination of high n, *rel*, and mid-level PC.	--
Simple Estimation Bias		
Standard IV	Overall, a negative and slightly bigger magnitude bias.	Small negative bias (overall and most sublevels).
MMB Full Constraint	Overall, a positive and lower magnitude bias; Much higher bias magnitude with low n, low PC, and unequal *var*	Small negative bias, most biased (overall and most sublevels).
MMB No Constraint[a]	--	Small negative bias, least biased (overall and most sublevels).
Relative Estimation Bias		
Standard IV	Same as Simple Estimation Bias	
MMB Full Constraint		
MMB No Constraint[a]		
Simple SE Bias		
Standard IV	Overall, a negative and smaller magnitude in bias. Within sublevels, mostly negative and mostly smaller magnitude in bias, except when PC=0.5 or 0.8. More sucetable to combination of n=50 or 100 and PC=0.1	Small positive bias, least biased (overall and sublevels).
MMB Full Constraint	Overall, a negative and bigger magnitude in bias. Within sublevels, mostly negative and mostly bigger magnitude in bias, except when PC=0.5 or 0.8. Less sucetable to combination of n=50 or 100 and PC=0.1	Small positive bias, second most biased (overall and sublevels, except when *var*=0.5).
MMB No Constraint[a]	--	Small positive bias, most biased (overall and sublevels, except when *var*=0.5).

Table 33 (Continued)

	Approach Comparison	
	All Conditoins	**Applicable Conditions**
Relative SE Bias		
Standard IV	Overall, very small magnitude negative bias and much smaller magnitude in bias. Within sublevels, mostly negative and mostly much smaller magnitude in bias, except when *PC*=0.5 or 0.8. Less sucetable to combination of *n* =50 or 100 and *PC*=0.1	Small positive bias, least biased (overall and sublevels).
MMB Full Constraint	Overall, larger magnitude negative bias and much bigger magnitude in bias. Within sublevels, all negative and mostly much bigger magnitude in bias, except when *PC*=0.5 or 0.8. More sucetable to combination of *n* =50 or 100 and *PC*=0.1	Small positive bias, second most biased (overall and sublevels, except when *var*=0.5).
MMB No Constraint[a]		Small positive bias, most biased (overall and sublevels, except when *var*=0.5).
Power		
Standard IV	Lower power (overall, across sublevels, and across combinations of n and PC). Difference disappear with PC=0.8.	Lowest (overall and across most sublevels)
MMB Full Constraint	Higher power (overall, across sublevels, and across combinations of n and PC). Difference disappear with PC=0.8.	Middle overall; highest for some sublevels.
MMB No Constraint[a]	--	Higest overall; highest for some sublevels.
Type I Error		
Standard IV	Closer to 0.05(overall and across sublevels).	Slightly smaller than MMB-NC
MMB Full Constraint	Much higher type I error rate (overall and across sublevels). Difference disappear with *PC*=0.8.	Closest to 0.05
MMB No Constraint[a]	--	Highest

Note. -- Not exist or not applicable. [a]Only applicable conditions were used for this dependent variable.

Across the six dependent variables, *n* and *PC* were the two most influential factors for the SIV and the MMB-FC approaches, especially with respect to the Success Indicator, Power, and the two SE bias measures. The two estimation bias measures were not greatly affected by any simulation factor. Only the MMB-FC method was slightly influenced by the interaction term *PC***var*. Apart from *n* and *PC*, other factors exhibitted their influence only on specific dependent variables. Measurement Reliability was very influential for estimation success, and Effect Size was important for empirical power. Note that it does not mean that terms not selected do not have effects on the dependent variables. In fact, most terms are influential, but for a researcher who is having trouble deciding where to allocate limited resources in order to have the best return, the factors selected here have more practical influence than those not selected. This simulation study chooses factor levels that are realistic in real studies. Even if expanding the ranges of some factors can make them look influential, the non-realistic levels will not provide enough practical guidance for researchers.

For the SIV and the MMB-FC approaches, the effects of the simulation factors are mostly as expected: favorable conditions lead to better estimation. However, there are several exceptions worth mentioning.

1) When the three latent classes have different Level 2 Covariance matrices, using the MMB Full Constraint estimation leads to more parameter estimation bias. When noncompliers have a smaller covariance matrix than the compliers, γ_c is underestimated. When noncompliers have a bigger covariance matrix, γ_c is overestimated.

2) In terms of the SE bias measures, for both estimation approaches, the combination of low *n* and low *PC* leads to severe underestimation of the SE. However, either increasing *n* or *PC* will yield a much more accurate estimation.
3) When *PC* is low, increasing *n* would not necessarily lead to a higher power if the MMB Full Constraint method is used.

As for the MMB-NC approach, only the first dependent variable was analyzed with all simulation conditions. Similar to the other two, *n* and *rel* were two influential factors, but *PC* was only influential for this estimation approach. The effect of *PC* is also worth mentioning: with *PC* = 0.5, the success rate reached its highest average. Due to the prevalent estimation non-convergence while using the MMB-NC approach, only very high *n* levels and not so extreme *PC* levels were used for the analyses of the last five dependent variables. With the limited conditions, the results show that the factors did not explain much variation in the dependent variables. The only exception was for the empirical power where *d* and *PC* both had sizable positive influences.

Comparing the three estimation approaches, the MMB-NC approach is out of discussion first because of its extremely low convergence rate under low *n* and low *PC*. The SIV approach should be the best choice considering all aspects. It has the highest overall success rate, the lowest estimation bias magnitude, and the lowest SE bias magnitude. Although the MMB Full Constraint method has much higher power, indicating higher empirical power, it is mainly due to the fact that this method on average overestimates γ_c and underestimates its standard error. This is evident with one unexpected finding: when *PC* is low, increasing *n* would not necessarily lead to a

higher power if the MMB Full Constraint method is used. When *PC* and *n* are both very low, γ_c is largely overestimated and its standard error is largely underestimated; hence, the significance rate was incorrectly inflated. As *n* becomes larger, the bias on both measures starts to shrink; therefore, the incorrectly inflated power also becomes closer to the truth. In practice, using this approach is likely to incorrectly detect a nonexistent treatment effect and therefore leads to a type I error. In fact, the MMB Full Constraint method did yield a much higher type I error rate than the other two methods.

Even if a researcher has the luxury to have a large sample size ($n \geq 500$) and a reasonable complier proportion ($0.3 \leq PC$), the SIV approach and the MMB-NC approaches were two very compatible choices. The former slightly underestimates the parameter more, while the latter slightly overestimates the SE more. As a result, the two approaches yields very similar power and type I error rate and both approaches do not inflate or deflate the power or type I error rate incorrectly. The MMB Full Constraint method, on the other hand, remains to be a flawed choice due to its largely inaccurate estimation of γ_c, especially with unfavorable conditions.

5.2.Summary of Practical Guidance on Sample Size and Measurement Reliability

The current study also strives to provide guidance on how to choose *n* and *rel*, the two manipulable factors for most experiments, to meet certain criteria on success rate, estimation bias, power, and type I error rate. Therefore, for each of the four dependent variables (i.e., Success Rate, Relative Estimation Bias, Relative SE Bias, and Significance Indicator), there is one table using cross tabulation to summarize the dependent variable with respect to *n*, *rel*, *PC*, and any other ANOVA analyses

selected factors. *PC* is consistently included because the selection of usable cells for the MMB-NC method depends on this factor.

Unfortunately, by using the selected factors together with any combination of *n*, *rel*, and *PC*, there is no guarantee, at least considering all simulation conditions used in this study, on all five measures. Specifically, none of the investigated configurations would lead to a 100% guarantee on type I error rate. With more preferable conditions, the probability of meeting the criteria is higher, but using *n*, *rel*, and *PC* alone, which are the selected factors for the dependent variable Significance Indicator, does not assure a negligible type I error rate.

There is still merit in comparing the conditions that guarantee a 100% satisfactory rate (within the current simulation design) on other measures. Firstly, by mapping conditions that lead to satisfactory success rate, negligible estimation bias, and negligible SE bias together, it is at least a high probability assurance for researchers and practitioners to obtain a converged and accurate estimation. Note that only *n*, *rel*, and *PC* were used for the three measures in this study, so the following discussion is limited to the three factors, too. The results show that the SIV approach is the most lenient on acceptable conditions. With *PC* = 0.8, the SIV approach yielded acceptable results on all three criteria even with a sample size of 200 regardless of measurement reliability. If *PC* was as low as 0.5, only the combination of *n* = 1,000 and *rel* = 0.8 satisfied all three measures. The MMB-NC approach and the MMB Full Constraint approach were much more limiting on the conditions. The former would only work with *n* = 1,000, *rel* = 0.8, and *PC* = 0.5 or 0.8, while the

latter would only work with *PC* = 0.8, *n* = 1,000, *rel* = 0.5 or 0.8 or *PC* = 0.8, *n* = 500, and *rel* = 0.8.

Second, after mapping empirical power together with the three criteria, the conditions became even more limiting. The following discussion is limited to the three factors above plus *d*. None of the three estimation approaches would guarantee satisfactory empirical power when *d* = 0.2. The SIV approach was still the most permissive approach among three. The smallest sample size with *PC* = 0.8 now increased to 500 with *d* = 0.8 for both measurement reliability values or with *d* = 0.5 with only *rel* = 0.8. When *PC* = 0.5, only *n* = 1,000, *rel* = 0.8, and *d* = 0.8 could guarantee a satisfactory empirical power. For the MMB-NC approach, with *n* = 1,000, *rel* = 0.8, and *PC* = 0.5, only *d* = 0.8 would lead to a satisfactory empirical power, and with *n* = 1,000, *rel* = 0.8 and *PC* = 0.8, both *d* = 0.5 and *d* = 0.8 would work. For the MMB Full Constraint method, with *PC* = 0.8, *n* = 1,000, *rel* = 0.5 or 0.8 or with *PC* = 0.8, *n* = 500, and *rel* = 0.8, both *d* = 0.5 and 0.8 could yield a satisfactory empirical power.

5.3.Limitations

The results of the current study provide insights to researchers and practitioners. They also present some challenges for further research. The results are based on a latent growth model where only four time points were used and a linear growth was implemented. Previous studies have shown that additional measurements, especially additional measurements that could increase the over construct reliability (Hancock & Mueller, 2001) might decrease estimation bias, increase convergence and increase statistical power (Tolvanen, 2007; Hancock & Gagné2006). In addition,

latent growth models are also widely used for studies where nonlinear growth trajectories are involved. It is likely that the current findings can be applied to LGMs with nonlinear trajectories, but more exploration regarding this issue can be worthwhile.

The exploration of the MMB-NC approach is somewhat limited due to the prevalent estimation non-convergence with regard to unfavorable sample size and complier proportion designs. Although the study provides useful suggestions for applied studies in terms of the model convergence, in order to probe more into the conditions yielding low convergence rate with respect to criteria (e.g., power, estimation bias), future studies can use some modification of the model estimation process.

Useful approaches include raising the number of random starting values sets, increasing replication times until certain convergence rate is met, and using real parameter values as starting values. Researchers should also be careful that these approaches also have their drawbacks. The first two methods means that the computation time can be very lengthy. The current study used 500 sets of random starting values. Although parallel computing was used to shorten simulation time, the time for 10% of the total simulation amount was more than three days with a powerful server. Raising the number of random starting value sets and increasing the replication times will definitely add more simulation time. In addition, some design cells have extremely low convergences rates (a lot of 0% convergence in fact), simply increasing the number of \starting values or replication time may not necessarily solve

the problem. It is almost safe to conclude that the MMB-NC approach is a too complex model for research settings with low sample size or low complier proportion.

Instead of trying to reach the adequate convergence rate while using small sample size designs, the more important question is to find the minimum sample size where "enough" number of datasets are converged and evaluation on other measures can proceed. The current study used 30% as the "enough" cutoff and classified $n \geq 500$ and $PC \geq 0.3$ as applicable conditions. However, there is a gap between the next available sample size, 200, and the cutoff sample size, 500. Incorporating more *n* levels between 200 and 500 would reveal more information on the behavior of the MMB-NC approach.

One goal of the present study is to compare the SIV method and the MMB method under broad conditions. The MMB method splits into two variants, one with full constraint and one with no constraint. While the no constraint approach cannot be applied to low sample size conditions, the full constraint approach is the only choice for the MMB method under such conditions. As covariance difference (*var*) is one design factor, the full constraint model is wrong for designs with $var \neq 1$. The comparison of the SIV and the MMB-FC approaches did not specifically distinguish different *var* levels, so it is unclear if the SIV approach will still outperform the MMB-FC if only $var = 1$ is used. As it is generally not recommended to have full constraint across classes (Little, 2013), and it is not the main focus for the current study, this study did not delve more into this area. Future study can be done by focusing on this topic. It is possible that if the three classes are more different the

MMB-NC method will be a more defendable choice. Future studies may pay more attention to this.

Compliance in this study was defined very clearly as taking treatment or not taking treatment. In reality, however, it is not always so well-defined. Low compliance can mean that participants partially follow the instruction of the experimental design. In other words, different participants may take different levels of dosage of the treatment. This issue is common in educational experiments, but the traditional CACE may not be a good approach for this scenario. A full mixture model can be a better solution here.

In this study, participant attrition was not included in the simulation design. However, attrition is a common problem for longitudinal studies. It is unclear how the estimation of CACE would be affected by attrition. On one hand, attrition can cause missing values and leads to more estimation biased and less estimation precision. More investigation on missing value could help to understand the influence of attrition. On the other hand, attrition can be also caused by the change of compliance membership. For example, a complier can become a non-complier as a study progresses. In this case, using models that can accommodate compliance class changing would be a preferable solution. More investigation regarding attrition should be conducted.

5.4.Implication for Future Studies

This study expands the literature of the longitudinal CACE estimation while using the LGM framework. The present study provides important evidence about how different research conditions can affect the estimation of the longitudinal CACE in

terms of the estimation success rate, estimation bias, power, and type I error rate. Further, this study provides guidance on choosing sample size and measurement reliability for researchers and practitioners. Last but not least, the study compares two estimation methods and demonstrates that the Standard IV approach is an overall better selection. These findings are important because the conditions included in the present study are realistic conditions for applied studies and the two estimation methods investigated are widely used in applied studies. With the guidance, researchers and practitioners can make a more educated decision for their research designs. Despite the limitations, this study compliments existing literature and provides a good starting point for future investigation.

Bibliography

Altonji, J. G., Elder, T. E., & Taber, C. R. (2005). An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. *Journal of Human Resources*, *40*, 791-821.

Angrist, J. D., & Imbens, G. W. (1995). Identification and estimation of local average treatment effects. *Econometrica*, *62*, 467–476

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*, 444-455.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*, 979-1041.

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

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2012). Causality and endogeneity: Problems and solutions. In D.V. Day (Ed.), *The Oxford handbook of leadership and organizations* (pp. 93-117). New York, NY: Oxford University Press.

Arbuckle, J. L. (2006). *Amos 7.0 User's Guide*. Chicago: SPSS.

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92-108). New York, NY: Guilford Press.

Bayley, N. (1956). Individual patterns of development. *Child Development*, *27*, 45-74.

Bell, R. Q. (1953). Convergence: An accelerated longitudinal approach. *Child Development*, *24*, 145-152.

Bell, R. Q. (1954). An experimental test of the accelerated longitudinal approach. *Child Development*, *25*, 281-286.

Bentler, P. M. (2006). *EQS 6 structural equation program manual*. Encino, CA: Multivariate Software, Inc.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8,* 225–246.

Bolton, R. N., & Drew, J. H. (1991). A longitudinal analysis of the impact of service changes on customer attitudes. *The Journal of Marketing*, *55*, 1-9.

Boruch, R., De Moya, D., & Snyder, B. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 50-79). Washington, DC: Brookings.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Brookhart, M. A., Rassen, J. A., & Schneeweiss, S. (2010). Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiology and Drug Safety*, *19*, 537-554.

Brookhart, M. A., Wang, P., Solomon, D. H., & Schneeweiss, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology*, *17*, 268-275.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis*. Newbury Park, C: Sage.

Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science*, *6*, 42-57.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107-112.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Dee, T. S. (2004). Are there civic returns to education? *Journal of Public Economics*, *88*, 1697-1720.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, *39*, 1-38.

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2013). *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Mahwah, NY: Lawrence Erlbaum Associates, Inc.

Edwards, M. C., Wirth, R.J., Houts, C. R., & Xi, N. (2012). Categorical data in the structural equation modeling framework. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 195-208). New York, NY: Guilford Press.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, *68*, 589-599.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1* (pp. 59-82). Berkeley, CA: University of California Press.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3-19.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509-529.

Farrington, D. P., Loeber, R., & Welsh, B. (2010). Longitudinal-experimental studies. In A. R. Piquero & D. Weisburd (Eds.), *Handbook of quantitative criminology* (pp. 503–518). New York, NY: Springer.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D., & Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *Journal of the American Statistical Association*, *99*, 239–249.

Frumento, P., Mealli, F., Pacini, B., & Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, *107*, 450-466.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*, 65-83.

Gao, X., Brown, G. K., & Elliott, M. R. (2014) Joint modeling compliance and outcome for causal analysis in longitudinal studies. *Statistics in Medicine*, *33*, 3453-3456.

Glewwe, P., Jacoby, H. G., & King, E. M. (2001). Early childhood nutrition and academic achievement: a longitudinal analysis. *Journal of Public Economics*, *81*, 345-368.

Hancock, G. R., Harring, J. R., & Lawrence, F. R. (2013). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 309-341). Greenwich, CT: Information Age Publishing, Inc.

Harris, D. N., & Goldrick-Rab, S. (2012). Improving the productivity of education experiments: Lessons from a randomized study of need-based financial aid. *Education*, *7*, 143-169.

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one-and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, *17*, 315-339.

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*, *38*, 214-216.

Heckman, J. J. (2008). Econometric causality. *International Statistical Review*,*76*, 1-27.

Herman, R. (1999). *An educator's guide to school wide reform*. Arlington, VA: Educational Research Service.

Hirano, K., Imbens, G. W., Rubin, D. B., & Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, *1*, 69-88.

Hogan, J. W., & Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, *13*, 17-48.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*, 945-960.

Hoyle, R. H. (2012). Introduction and overview. In R. H. Hoyle (Eds.), *Handbook of structural equation modeling* (pp. 3-16). New York, NY: Guilford Press.

H.R. 3801. Education Sciences Reform Act of 2002 (107th Congress). Retrieved September 16, 2003, from http://thomas.loc.gov

Hsieh, P., Hsieh, Y.P., Chung, W.H., Acee, T., Thomas, G.D., Kim, H.J., You, J., Levin, J.R., and Robinson, D.H. (2005). Is educational intervention research on the recline? *Journal of Educational Psychology*, *97*, 523-529.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1* (pp. 221–233). Berkeley, CA: University of California Press.

Ialongo, N. S., Werthamer, L., Kellam, S. G., Brown, C. H., Wang, S., & Lin, Y. (1999). Proximal impact of two first-grade preventive interventions on the

early risk behaviors for later substance abuse, depression, and antisocial behavior. *American Journal of Community Psychology*, *27*, 599-641.

Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statistical Science*, *29*, 323-358.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics, 25,* 305-327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies, 64,* 555-574.

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *62*, 257-270.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, *27*, 385-409.

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods*, *13*, 1-18.

Jo, B., & Muthén, B. O. (2001). Modeling of intervention effects with noncompliance: A latent variable approach for randomized trials. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in*

structural equation modeling (pp. 57-87). Mahwah, NJ: Lawrence Erlbaum Associates.

Jo, B., & Muthén, B. O. (2003). Longitudinal studies with intervention and noncompliance: Estimation of causal effects in growth mixture modeling. In S. P. Reise and N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 112-139). New York, NY: Lawrence Erlbaum Associates, Inc.

Jöreskog, K. G., & Sörbom, D. (20015). *LISREL (version 9.2) [computer software]*.Chicago, IL: Scientific Software International.

Kang, S. H., & Sivaramakrishnan, K. (1995). Issues in testing earnings management and an instrumental variable approach. *Journal of Accounting Research*, *33*, 353-367

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111-125). New York, NY: Guilford Press.

Lin, J. Y., Have, T. R., & Elliott, M. R. (2009). Nested Markov compliance class model in the presence of time-varying noncompliance. *Biometrics*, *65*, 505-513.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Little, T. D., & Rhemtulla, M. (2013). Planned missing data designs for developmental researchers. *Child Development Perspectives*, *7*, 199-204.

Little, R. J., & Yau, L. H. (1998). Statistical techniques for analyzing data from prevention trials: treatment of no-shows using Rubin's causal model. *Psychological Methods*, *3*, 147.

Lockwood, C. M., & MacKinnon, D. P. (1998, March). Bootstrapping the standard error of the mediated effect. In *Proceedings of the 23rd Annual Meeting of SAS Users Group International* (pp. 997-1002). Cary, NC: SAS Institute, Inc.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford: Clarendon Press.

McArdle, J. J. (1989). A structural modeling experiment with multiple growth functions. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota symposium on learning and individual differences* (pp. 71-117). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McArdle, J. J., & Bell, R. Q. (2000). An introduction to latent growth models for developmental data analysis. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 69-107). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*, 110-133.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Eds.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

Miles, J., & Shevlin, M. (2001). Applying regression and correlation: A guide for students and researchers. London: Sage.

Mosteller, F., & Boruch, R. F. (Eds.). (2002). *Evidence matters: Randomized trials in education research*. Washington, DC: Brookings.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. New York, NY: Oxford University Press.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*, 81-117.

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143-165). Boca Raton, FL: CRC Press.

Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, *3*, 371−402.

Muthén, L. K., & Muthén, B. O. (1998-2018). *Mplus user's guide* (8thed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599-620.

Nevitt, J., & Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Structural Equation Modeling*, *8*, 353-377.

Poi, B. P. (2004). From the help desk: Some bootstrapping techniques. *Stata Journal*, *4*, 312-328.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*, *26*, 195-239.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*, 726-748.

Rogosa, D. R., & Willett, J. B. (1985). Understanding correlates of change by modeling individual differences in growth. *Psychometrika*, *50*, 203-228.

Rosseel, Y. (2012). lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software*, *48(2)*, 1-36.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics*, *2*, 881-935.

RStudio (2009-2017). RStudio: Version 1.1.383. Boston, MA: RStudio, Inc.

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

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization tets," by D. Basu. *Journal of the American Statistical Association*, *75*, 591-593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472-480.

Salkind, N. J. (2010). *Statistics for people who (think they) hate statistics: Excel 2010 Edition*. New York, NY: Sage.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334-1344.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

Shipley, B. (2002). *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference*. Cambridge: Cambridge University Press.

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*, 137-167.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21.

Sobel, M. E. (1995). Causal inference in the social and behavioral sciences. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.) *Handbook of statistical modeling for the social and behavioral sciences* (pp. 1-38). New York, Ny: Springer US.

Stoolmiller, M., Duncan, T., Bank, L., & Patterson, G. R. (1993). Some problems and solutions in the study of change: significant patterns in client resistance. *Journal of Consulting and Clinical Psychology*, *61*, 920-928.

Sussman, J. B., & Hayward, R. A. (2010). An IV for the RCT: Using instrumental variables to adjust for treatment contamination in randomized controlled trials. *Bmj*, *340*, c2073.

Thompson, M. S., & Green, S. B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.) (pp. 163-218). Greenwich, CT: Information Age Publishing, Inc.

Tolvanen, A. (2000). *Latenttien kasvukäyrä- ja simplex-mallien teoriaa ja sovelluksia pitkittäisaineistoissa kehityksen ja muutoksen analysointiin*. Licentiate thesis, Department of Statistics, University of Jyväskylä, Finland.

Tolvanen, A. (2007). *Latent growth mixture modeling: A simulation study*. Doctoral dissertation, Department of Mathematics, University of Jyvaskyla, Finland.

Vinokur, A. D., Price, R. H., & Schul, Y. (1995). Impact of JOBS intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, *23*, 39-74.

White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, *48*, 817–838.

Whitehurst, G. J. (2008). *Rigor and Relevance Redux: Director's Biennial Report to Congress*. IES 2009-6010. Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Wille, B., Beyers, W., & De Fruyt, F. (2012). A transactional approach to person-environment fit: Reciprocal relations between personality development and

career role growth across young to middle adulthood. *Journal of Vocational Behavior*, *81*, 307-321.

Wooldridge, J. (2012). *Introductory econometrics: A modern approach*. Mason, OH: Cengage Learning.

Wright, P. G. (1928). *Tariff on animal and vegetable oils.* New York, NY: Macmillan.

Wright, S. (1918). On the nature of size factors. *Genetics*, *3*, 367-374.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics*, *5*, 161-215.

Yau, L. H., & Little, R. J. (2001). Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, *96*, 1232-1244.