ABSTRACT

Title of dissertation: MULTILEVEL REGRESSION DISCONTINUITY MODELS WITH LATENT VARIABLES

Monica Morell, Doctor of Philosophy, 2020

Dissertation directed by: Professor Ji Seung Yang
Department of Human Development and Quantitative Methodology

Regression discontinuity (RD) designs allow estimating a local average treatment effect (LATE) when assignment of an individual to treatment is determined by their location on a running variable in relation to a cutoff value. The design is especially useful in education settings, where ethical concerns can forestall the use of randomization. Applications of RD in education research typically share two characteristics, which can make the use of the conventional RD model inappropriate: 1) The use of latent constructs, and 2) The hierarchical structure of the data. The running variables often used in education research represent latent constructs (e.g., math ability), which are measured by observed indicators such as categorical item responses. While the use of a latent variable model to account for the relationships among item responses and the latent construct is the preferred approach, conventional RD analyses continue to use observed scores, which can result in invalid or less informative conclusions. The current study proposes a multilevel latent RD model which accounts for the prevalence of clustered data and latent constructs in education research, allows for the generalizability of the LATE to individuals further from the cutoff,

and allows researchers to quantify the heterogeneity in the treatment effect due to measurement error in the observed running variable. Models are derived for two of the most commonly used multilevel RD designs. Due to the complex and high-dimensional nature of the proposed models, they are estimated in one stage using full-information likelihood via the Metropolis-Hastings Robbins-Monro algorithm. The results of two simulation studies, under varying sample size and test length conditions, indicate the models perform well when using the full sample with at least moderate-length assessments. A proposed model is used to examine the effects of receiving an English language learner designation on science achievement using the Early Childhood Longitudinal Study. Implications of the results of these studies and future directions for the research are discussed.

MULTILEVEL REGRESSION DISCONTINUITY MODELS WITH LATENT VARIABLES

by

Monica Morell

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
2020

Advisory Committee:
Professor Ji Seung Yang, Chair/Advisor
Professor Yang Liu, Co-Chair
Professor Gregory Hancock
Professor Eric Slud
Professor Laura Stapleton

Dedication

Para mis abuelas.

Acknowledgments

I would first like to thank my advisor, Dr. Ji Seung Yang, whose mentoring, encouragement, and guidance helped me achieve more than I had thought possible. What I have learned from her throughout my graduate career has been invaluable. I would also like to thank my dissertation committee members: Dr. Yang Liu for his guidance in this and other research; Drs. Gregory Hancock and Laura Stapleton for their mentoring and encouragement; Dr. Eric Slud for his patience and insights. A special thank you to Dr. Allan Wigfield for his support and to Dr. Peter Steiner for lending his expertise.

I also want to express my appreciation for the entire Educational Measurement and Statistics program at the University of Maryland, with its culture of collaboration, respect, and support. I am especially grateful for the guidance and encouragement Dr. Jeff Harring has provided me. I would also like to give a very heartfelt thank you to my peers, both current and former, in the program. The intellectual and emotional support they have given me has been inestimable.

Finally, I would like to thank both my parents, my brother Gaby, and Evan. Without their love, constant encouragement, and unwavering support I would not be where I am today.

Table of Contents

Chapter 1: Introduction

With the passage of the Every Student Succeeds Act (ESSA) in 2015, a reaffirming of the importance of quality education research, there has been a resurgence of interest in ensuring education research can support valid causal claims. The ability to draw causal conclusions from research is largely based on the research design used. To this end, designs that randomly assign participants to treatment groups, such as randomized control trials (RCTs), are considered the gold standard. However, the random assignment of students, teachers, or schools to interventions can be infeasible or unethical. For this reason, the regression discontinuity (RD; Thistlethwaite & Campbell, 1960) design is often used in education research, as it allows for causal conclusions, under strong model assumptions, without the need for randomization.

In the RD design, individuals are assigned to treatment conditions based on whether their scores on a running variable (also called an "assignment variable" or a "forcing variable") are above/below a specified cutoff value. Running variables (RVs) can be any continuous variable on which all participants have a non-missing value (e.g., age, pre-test scores, language proficiency). For example, when students are placed in a remedial math course based on whether their scores on a placement exam meets a given threshold, an RD study may be conducted to estimate the causal effect of the remedial math course on an outcome of interest. The causal treatment effect estimate provided by this design is generalizable to the subpopulation with RV scores at the cutoff (see Section 2.2), and is referred to as the local average treatment effect (LATE).

As there are many existing education interventions and programs that target at-need groups and, therefore, have existing eligibility criteria with a threshold for receiving the intervention, there are many opportunities to apply the RD design in education research. However, two other common features of education data can make the use of the conventional RD model challenging: 1) data collected in education settings tend to be clustered (e.g., students nested in classrooms, classrooms nested in schools), and 2) the variables of interest are often latent constructs that cannot be directly measured. The first feature poses a challenge as clustered data violate a key assumption of the conventional RD model, independence of observations, which can lead to invalid conclusions (see Section 2.4). The second feature requires researchers to decide how to calculate scores for the latent constructs before using them in the RD model. This decision can impact the quality of the RD treatment estimates (see Section 2.3).

As latent constructs cannot be directly measured, instead, data are collected on observed indicators (e.g., item responses on an exam). These item responses can be averaged or summed to calculate an observed score for the latent construct. However, these scores will contain measurement error. When used as predictors in a regression model, observed scores can attenuate regression coefficients resulting in invalid conclusions. A preferred approach is to calculate latent variable (LV) scores through the use of an LV model, which takes into account the unique relation between each item response and the latent construct, and consequently accounts for the measurement error. While there has been a trend in many fields towards extending statistical models to include LVs, RD continues to treat latent constructs as observed variables.

The tradition of using observed variables in RD stems from the belief that the measurement error in the observed RV (ORV) acts to make the RD design a local RCT and further allows researchers to examine the heterogeneity in the treatment effect and to generalize the treatment effect away from the cutoff value (D. S. Lee & Lemieux, 2010). Moreover,

when the interpretation of the treatment effect is confined to the scale of the ORV, any attenuation in the regression parameters due to measurement error in the ORV is not a concern (see Section 2.2.6). However, when researchers are interested in interpreting the treatment effect with respect to the latent RV (LRV), the results of a conventional RD analysis may be invalid, e.g., parameter estimates may be biased. Moreover, the conventional RD model is unlikely to allow for adequate generalizability of the LATE or quantification of the heterogeneity in the treatment effect without an explicit specification of a measurement model for the RV and additional assumptions (see Section 2.2.1).

Therefore, in order to account for the prevalence of clustered data and latent constructs in education research, as well as to increase the generalizability of the treatment effect to individuals further from the cutoff and allow researchers to quantify the heterogeneity in the treatment effect due to measurement error in the ORV, the present study proposes multilevel RD models in which the RV, covariates, and outcome variables are all latent. The proposed models are estimated using a reasonably efficient algorithm; parameter recovery is explored in two simulation studies. The recovery of the LATE under the proposed models is compared to the conventional approach in a simulation study; the proposed model is used in an empirical data application.

The rest of the chapter is structured as follows. First, I provide an overview of the limitations of the current approach to RD analysis in education settings using latent constructs. Next, I present the purpose of the current study and the specific study aims, which address these limitations. Finally, I discuss the possible contributions of this study to education researchers. The chapter concludes with an overview of subsequent chapters.

1.1 Statement of the Problem

The conventional RD approach fails to properly account for the hierarchical nature of education data and the use of latent constructs as predictors. This may not only result in invalid statistical conclusions (e.g., increased Type I error and bias in parameter estimates), but also limits the researcher's ability to gain more information about the effect of the treatment being studied. In an RD design, participants are assessed on a continuous RV and assigned to treatment conditions based on their scores on that variable in relation to an exogenous cutoff point. Participants with scores on one side of the cutoff point are assigned to the treatment group and those on the other side receive the comparison condition. After participants are assigned and the treatment is administered, they are assessed on an outcome measure. A regression is fit to predict outcome scores based on the RV and a dummy variable that indicates whether the participant received treatment or not. The effect of the treatment is estimated by comparing the regression lines of the two groups. Treatment effect estimates at the cutoff are unbiased when assumptions are met (Goldberger, 1972; Rubin, 1977). See Section 2.2.1 for more details on RD model assumptions.

Data and the fitted regression lines from a hypothetical RD study are depicted in Figure 1.1. The vertical line at an RV score of -0.5 designates an arbitrary cut-point above which participants are assigned to treatment and below which they are assigned to the control group. The treatment in this example has a positive effect, as depicted by the sharp increase in outcome scores at the cutoff (or discontinuity).

Opportunities for RD often emerge naturally in education as many programs or interventions targeting individuals or groups have eligibility cutoffs built into them. Moreover, because the RV can be unrelated to the outcome variable (Cook, Campbell, & Shadish,

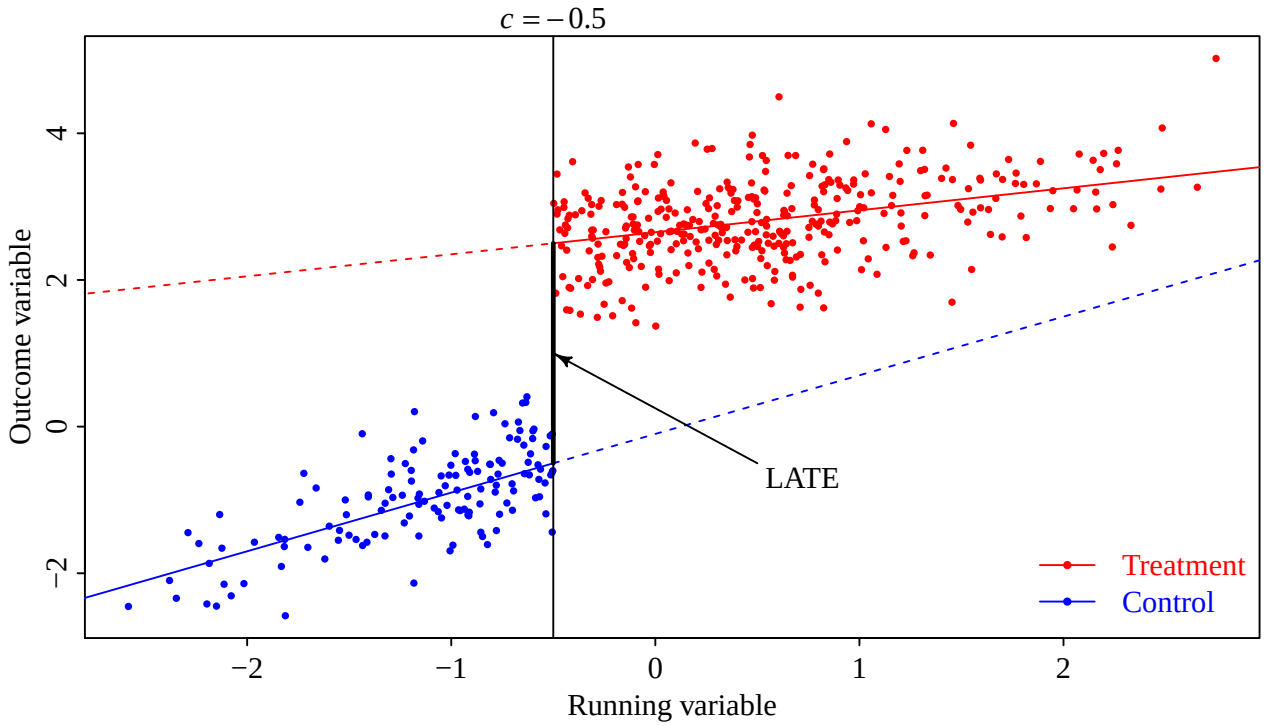

Figure 1.1: Graphical representation of a regression discontinuity model with an interaction term with a cutscore at -0.5. Participants with running variable scores above the cutoff are assigned to the treatment group; participants with running variable scores below the cutoff are assigned to the control group. The treatment has a positive effect as indicated by the local average treatment effect (LATE).

2002), RD maximizes a program or intervention's ability to use high-quality assessments in identifying individuals with the greatest need and assigning them to the appropriate treatment condition (Linden, Adams, & Roberts, 2006). Because the RD design complements existing features of education interventions, many such studies exist. However, the conventional RD framework fails to account for two features common in education applications: the multilevel nature of the data and measurement error in the variables.

In education research, RVs are often latent constructs (e.g., math proficiency), which are not directly observable. Instead, data are collected on response variables such as continuous or categorical item responses. Observed variable scores are calculated using the item-level raw scores by summing or taking an average (i.e., total number of items correct on an

exam, average of Likert scale responses, etc.). These noisy measures are then used as RVs in RD analyses. For instance, several RD studies have used school assessment scores (Chay, McEwan, & Urquiola, 2005; Chiang, 2009; B. A. Jacob & Lefgren, 2004a), entrance exam scores (Ding & Lehrer, 2007), standardized test scores (B. A. Jacob & Lefgren, 2004b; Van der Klaauw, 2008) and subject specific assessment scores (Goodman, 2008; Matsudaira, 2008) as RVs. Alternatively, psychometricians often rely on LV measurement models, such as item response theory (IRT; see e.g., Lord, 1980, for a review) models, to account for the association between observed indicators and latent constructs. Each construct is quantified by an LV; the distribution of observed indicators is governed by the LV and measurement model parameters (e.g., factor loadings, item difficulties). When LVs are used in the formulation of structural models (e.g., a regression discontinuity model), the simultaneous estimation of the measurement and structural model parameters is considered the gold standard (see e.g., Bartholomew & Knott, 1999; Bollen, 1989; Muthén, 2002; Rabe-Hesketh & Skrondal, 2004). Despite the trend of LV modeling in the social sciences, RD continues to favor an observed variable approach.

However, the issue of measurement error is more complex in the RD design. While it can be seen as having a positive effect on making the RD a localized RCT at the cutoff, it also limits the information researchers can get from the treatment effect. In a conventional RD analysis, the RV, outcome, and covariates are all treated as observed variables. Measurement error in the ORV is purported to have a positive role in converting an RD design to a local randomized control trial and allows researchers to examine heterogeneity and generalizability of the LATE (D. S. Lee & Lemieux, 2010). However, those aspects of the LATE are unlikely to be revealed by a conventional RD analysis without an explicit specification of a measurement model for the RV. Furthermore, when the RV is observed, the associated LATE is only identified on the ORV scale in the unit of the observed outcome. When the interpretation of the LATE is intended with respect to the latent RV (LRV)

in the unit of the latent outcome, the conventional RD analysis may be misleading as the LATE with respect to the ORV and the LATE with respect to the LRV are not necessarily the same. Therefore, introducing LVs to the RD analysis proves to be beneficial whether researchers are interested in the heterogeneity and generalizability of the LATE on the ORV scale or inferences with respect to the LATE on the LRV scale.

Most education data are hierarchical in nature due to the structure into which the education system is organized: Students receive instruction in classrooms which exist inside schools, which are grouped by districts. This system makes it difficult to implement random assignment. For example, district officials may be uncomfortable with the unequal allocation of resources between schools when only some are assigned to a treatment, principals may be concerned with disruptions to regular school routines and deadlines, and parents may protest if their child is not randomly assigned to receive a potentially beneficial intervention (Cook, 2002). This hierarchical structure gives rise to multilevel data, which may create challenges for statistical modeling. Because observations are no longer independent, a key assumption underlying a majority of parametric procedures is violated making the use of many traditional statistical models inappropriate (Burstein, 1980; Goldstein, 2011; Kreft & De Leeuw, 1998; Raudenbush, 1993; Snijders & Bosker, 2012).

In an RD design, it may be the case that treatment assignment occurs at the cluster level; for instance, an entire school may qualify for participation in a governmental intervention program. Not accounting for such sampling structures can lead to improper inferences. On the one hand, ignoring the nested structure of data (e.g. Banks & Mazzonna, 2012; B. A. Jacob & Lefgren, 2004b; Matta, Ribas, Sampaio, & Sampaio, 2016; Ou, 2010), i.e., disaggregation, violates the usual regression assumption of independence of errors which may lead to underestimated standard errors. On the other hand, aggregation, analyzing the data at the cluster level instead of the individual observation level (e.g., Holbein & Ladd, 2017; Tang, Cook, Kisbu-Sakarya, Hock, & Chiang, 2017), is only valid if the treatment

effect is inferred at the cluster level. However, as Burstein (1980) noted, aggregation bias is especially salient in education settings because "aggregation inflates the estimated effects of student background on outcomes and decreases the likelihood of identifying lower level characteristics and practices that are effective."

More adequate modeling strategies include pooling the results of separate one-level RD analysis for each cluster (e.g., Eggers, Fowler, Hainmueller, Hall, & Snyder Jr, 2015; Gamse, Bloom, Kemple, & Jacob, 2008; Van der Klaauw, 2002), fitting multilevel RD models (e.g., Boatman & Long, 2010; Li, Mattei, & Mealli, 2015a; Tuckwiller, Pullen, & Coyne, 2010), and correcting the standard errors of the regression coefficients if cluster effects are not of interest (e.g., Cliffordson & Gustafsson, 2010; Sojourner, Frandsen, Town, Grabowski, & Chen, 2015). However, the integration of LV measurement models and multilevel RD analyses has not been explored in the existing literature.

In summary, the conventional application of the RD model in many education studies is inadequate as the model does not account for the hierarchical data or use of latent constructs. Moreover, the conventional model does take advantage of the ability to model the measurement error in the RV in order to improve the interpretability of the RD treatment effect estimates.

1.2 Purpose of the Study

Given the value of RD designs in causal inference, the hierarchical nature of education data, and the widespread popularity of LV modeling in social sciences research, the integration of RD analysis, multilevel modeling, and LV modeling would benefit education researchers who seek causal evidence by providing more information about the effect of the intervention. The purpose of the proposed study is to derive multilevel LV RD models in which the outcome, RV, and covariates are latent, and estimate the proposed models with

a reasonably efficient algorithm. The performance of the LV RD analyses will be assessed and compared to the performance the conventional RD approach, i.e., using ORVs. The research questions are as follows:

1. How well does the developed code recover parameters of interest in the proposed models, e.g., RD treatment effect, IRT item parameters, and variance components?

2. How are the RD treatment effect estimates in the proposed models recovered under different simulation conditions, e.g., sample size and test length?

3. How do the proposed models perform compared to the conventional observed-variable approach in simulation studies?

4. How does the proposed model perform when applied to empirical data?

The research hypotheses are:

1. The developed code will recover model parameters well under ideal conditions.

2. The RD treatment effect estimates will be well recovered under large sample sizes and long tests lengths when using the full sample.

3. The proposed models will recover the treatment effect estimate better than the conventional observed-variable approach under all conditions.

4. The proposed model will provide more comprehensive treatment effect estimates than the conventional model in an empirical data application.

1.3 Significance of the Study

There are three main areas in which the proposed work may make significant contributions. Firstly, the proposed model will allow researchers to gain more information about

the effect of the treatment being studied. The RD design is popular in policy research, partly because of the high internal validity of the LATE. However, a notable weakness of the design is the limited external validity of the LATE (i.e., the LATE applies only to the subpopulation with RV scores within a narrow window of the cutoff). The proposed work will improve upon this limitation by both increasing the generalizability of the LATE to participants further from the cutoff (see Section 2.2.4) and by allowing researchers to quantify how much the LATE differs across individuals with the same observed score (i.e., the heterogeneity in the LATE due to measurement error; see Section 2.2.5) under additional assumptions (see Section 2.3.2). Furthermore, researchers will be able to interpret the treatment effect with respect to the latent construct (i.e., the LRV) rather than just the ORV (see Section 2.2.6). As many RD studies in education use measures of latent constructs (e.g., depression, anxiety, math ability) as RVs, the proposed model can be applied widely in the education field so that stakeholders may draw more informed conclusions about the efficacy of the interventions being studied. For example, a study which aims to evaluate a gifted reading program using conventional RD analysis may use the score on a reading exam (e.g., number correct or percentage correct) as the RV. While this allows the researcher to interpret the LATE with respect to the observed score on the reading exam, which is applicable only to students with similar reading scores on that same exam, it does not allow for an interpretation of the LATE with respect to reading ability, the latent construct underlying the observed score. In order to interpret the LATE with respect to the latent construct, estimate the treatment effect for a larger subpopulation, or examine the heterogeneity in the LATE due to the measurement error in the ORV, a measurement model must be correctly specified for the RV. The proposed model includes such a specification and, therefore, increases the information that the RD analysis can provide about the treatment. Consequently, the proposed model may be used broadly as it not only allows for the treatment effect to be estimated with respect to the latent construct, which has broader

implications than its observed counterpart, but also improves inferences on the ORV scale by allowing for an an estimate of the treatment effect at ORV scores away from the cutoff as well as for an examination on how the LATE varies due to measurement error in the ORV.

The second area in which the proposed model may make contributions is in creating stronger scientific links among studies. When developing instruments and assessments, psychometricians use LV models, e.g., IRT models, which explicitly specify the relationships between item responses and the latent constructs. In calculating observed scores for these instruments (e.g., sum or average of item responses) researchers are ignoring the relationships assumed when the instruments were developed. Moreover, observed scores contain less information about the latent construct than LV scores. As such, consistently using LV scores from instrument conception through structural models is the more desirable approach as it maintains the assumptions and meanings intended when the instrument was developed. The proposed model will aid in this endeavor as it allows researchers to specify the appropriate measurement model for the latent constructs. Furthermore, using LV models will allow researchers greater flexibility when collecting data from multiple sites or across multiple time points as different forms of tests that measure the same latent construct can be used; the scores can later be linked onto a common scale (see e.g., Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Lim, 1993a, 1993b; Mislevy, 1992). Furthermore, when the RD analysis is intended to validate the effectiveness of an intervention program that has been previously evaluated, integrated data analysis can be performed by concatenating the data sets, as long as measurement invariance can be established for at least a subset of items (i.e., anchor items) and the fit of the measurement model is reasonable.

Finally, while many large datasets (including multi-site and multi-year data) measuring latent constructs on which RD analysis can be conducted are available for public or restricted use (e.g., Institute of Education Sciences, National Science Foundation), most

of these data include only observed variable scores for the latent constructs and do not report item-level responses. This not only encourages researchers to use observed variable scores as latent variable scores cannot be calculated without more information, it also deemphasizes the importance of working with high quality, reliable, and valid instruments as the provided data cannot be used to these ends. The proposed model requires the use of item-level data to provide the additional estimates of an intervention's effect (i.e., increased generalizability of the treatment effect and quantification of the heterogeneity in the treatment effect), the practice of reporting item-level data may become more commonplace. It is noted that recording item-level response data and validating measurement models will eventually contribute to the improvement of research practices and accumulations of scientific evidence.

1.4 Overview of the Chapters

Chapter 2 (Literature Review) introduces causal inference in education settings and provides an overview of the conventional RD design, including limitations in the interpretability of the LATE. The use of latent constructs in RD analyses is explored and the typical treatment of multilevel data in RD modeling is reviewed. Estimation approaches for multilevel LV models are then introduced and the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm and its general implementation are presented. In Chapter 3 (Multilevel RD Models with Latent Variables and Estimation), a general multilevel LV RD modeling model, as well as two common multilevel designs, are derived. The specific implementation of the MH-RM algorithm to estimation the proposed models is then presented. In Chapter 4 (Simulation Studies) two simulation study designs are presented and the results of the studies are described. Chapter 5 (Empirical Data Analysis) presents an empirical data set, describes the use of the LV RD model applied, and presents the results. Chapter 6

(Discussion) includes a summary of the study, a discussion of the results, and implications for future research.

Chapter 2: Literature Review

2.1 Causal Inference in Education Research

The importance of drawing causal conclusions about the effects of an intervention on student achievement was reaffirmed with the passage of the Every Student Succeeds Act (ESSA) in 2015. The legislation promotes the use of "evidence-based interventions" as the foundation of the education programs schools choose to implement. ESSA defines four tiers of evidence which correspond with research designs. Experimental studies provide "strong" evidence, while quasi-experimental studies yield "moderate" evidence and correlation studies provide "promising" evidence. The least desirable evidence is from research-based theory (S.1177-290, Section 8002). The ability to draw causal conclusions from studies is largely determined by the research design employed. Designs that randomly assign participants to treatment conditions, such as randomized control trials (RCTs), meet the ESSA's definition for strong evidence. Randomly assigning participants creates groups that are probabilistically similar to each other on average. The effect of an intervention can therefore be measured by the differences observed between the groups after the treatment is administered (Cook et al., 2002). In the potential outcomes framework (Holland, 1986; Rubin, 1974), an individual i, has two possible outcomes, $Y_i^{[1]}$ and $Y_i^{[0]}$, which would result from receiving the treatment condition and receiving the control condition, respectively. The causal effect of treatment on an outcome for individual i, is the difference between that individual's two potential outcomes, i.e., $(Y_i^{[1]} - Y_i^{[0]})$. The fundamental problem of causal

inference is that both outcomes cannot be observed for the same individual at the same time (Holland, 1986). Instead a participant only receives one treatment condition, and the counterfactual condition (i.e., the treatment condition the participant does not receive) cannot be observed. Therefore, the effect of treatment for an individual cannot be calculated. Instead, the average treatment effect (ATE) for a population can be calculated as the expected value of the individual treatment effects, or

$$ATE = E\{Y^{[1]} - Y^{[0]}\}, \tag{2.1}$$

which is the same as the difference between expected population outcomes under both conditions, or

$$ATE = E\{Y^{[1]}\} - E\{Y^{[0]}\}, \tag{2.2}$$

This calculation is possible because the counterfactual is estimated via the random selection and random assignment of participants (Murnane & Willett, 2010). However, some circumstances can mitigate against the use of random assignment (see e.g., Cook, 2002, for a review). For example, in education research, randomly assigning participants to treatment conditions may not be ethical (e.g., withholding compensatory education program from at need students) or be feasible (e.g., schools not willing to use different teaching programs across classrooms). In these cases, an average treatment effect (ATE) can be estimated via an alternative assignment mechanism, such as the one used in the RD design.

The RD design is the strongest method for estimating causal effects after RCTs (Cook, 2008; Cook et al., 2002), and the only non-randomized experimental design that has an unbiased and consistent estimator (Rosenbaum, 1995; Shadish, 2011). Its usefulness as an

alternative to RCTs is supported by several within-study comparisons which have shown RD methods yield unbiased estimates of treatment effects comparable to those in RCTs (Aiken, West, Schwalm, Carroll, & Hsiung, 1998; Berk, Barnes, Ahlman, & Kurtz, 2010).

2.2 The Regression Discontinuity Design

The RD design may be used in situations where participants are assigned to treatment or control based on whether their value on a continuous variable exceeds a predetermined cutoff value. To formalize the conventional RD model, let $i = 1, \ldots, n$ indicate participants in a sample of size n. For each participant i, let Y_i be the observed outcome, X_i be the ORV, which is measured prior to treatment, and $T_i \in \{0, 1\}$ be a deterministic function of the RV such that $(T_i = 0)$ indicates the participant received the control condition and $(T_i = 1)$ indicates the participant received treatment, based on where X_i falls in relation to a cutoff, c (e.g., $T_i = 1\{X_i < c\}$). Generally, the conventional RD model can be expressed as

$$Y_i = T_i f_1(X_i) + (1 - T_i) f_0(X_i) + \varepsilon_i, i = 1, \ldots, n,\ \varepsilon_i \sim (0, \sigma^2) \tag{2.3}$$

where $f_1(x)$ is the expected outcome when a participant with an ORV value $X_i = x$ receives treatment, and $f_0(x)$ is the expected outcome when the participant does not receive treatment. If both functions are linear, i.e., $f_1(x) = \alpha_0 + \alpha_1(x - c)$ and $f_0(x) = \beta_0 + \beta_1(x - c)$, then Equation 2.3 reduces to

$$Y_i = \beta_0 + \beta_1(X_i - c) + \beta_2 T_i + \beta_3(X_i - c)T_i + \varepsilon_i,\ \varepsilon \sim N(0, \sigma^2), \tag{2.4}$$

in which the intercept, β_0, is the expected outcome score for individuals in the control group with an ORV at the cutoff, β_1 is the relation between the ORV and the outcome, β_2 is the treatment effect, and β_3 is the interaction between the RV and the treatment effect.

A graphical illustration of Equation 2.4 appears in Figure 1.1. Recall that the vertical line at the RV score of -0.5 represents the cutoff value above which participants are assigned to treatment and below which they are assigned to the control group. A positive LATE is seen by the jump upwards at the cutoff; this is the conceptualization of RD analysis as a "discontinuity at a cutoff point" (Hahn, Todd, & Van der Klaauw, 1999), in which the magnitude and direct of the jump is a measure of the ATE locally at the cutoff. RD can also be characterized as local randomization (Cattaneo, Frandsen, & Titiunik, 2015; D. S. Lee, 2008): Participants in an infinitesimal neighborhood of the cutoff point are deemed identical on average, and consequently the difference observed in the outcome is attributed to treatment.

Definition of the Treatment Effect

In the conventional RD model (Equation 2.4), it can be seen that $E(Y_i^{[1]}|X_i = c) = \lim_{x \downarrow c} E(Y_i|x) = \beta_0 + \beta_2$ when the RV, X_i, approaches the cutoff c from above; similarly, $E(Y_i^{[0]}|X_i = c) = \lim_{x \uparrow c} E(Y_i|x) = \beta_0$ when the RV approaches c from below. Therefore, the coefficient

$$\beta_2 = \lim_{x \downarrow c} E(Y_i|x) - \lim_{x \uparrow c} E(Y_i|x) \tag{2.5}$$

amounts to the ATE within the "subpopulation" identified by $X_i = c$, which is referred to as the LATE (Hahn, Todd, & Van der Klaauw, 2001).

2.2.1 Assumptions of the RD Model

One critical assumption of the RD design is perfect adherence to the thresholding rule: That is, the probability of receiving treatment changes perfectly from 0 to 1 at the cut-off point (Hahn et al., 2001). This assumption can be checked via a visual inspection of a graph, such as that in Figure 1.1, for discontinuities at places other than the cut-point

and by formal hypothesis test (McCrary, 2008). Meeting this assumption requires perfect adherence to the treatment assignment for all participants, i.e. all participants must be compliers. This ideal case is often referred to as a "sharp" RD design. In contrast, when some participants are defiers, it is typically referred to as a "fuzzy" RD design. Fuzzy RD designs can occur when some treatment group members do not receive treatment while some control group members do (Bloom, 1984; Bloom et al., 1997). While RD analyses can still be performed under a fuzzy design by taking into account the effect of the defiers on the estimated treatment effect (Bloom, 2012), the current work focuses on sharp RD only. Note, the "sharp" label refers to the ORV only in both the conventional RD model and the proposed models. The second assumption is that the relation between the RV and the outcome variable is correctly specified (Hahn et al., 2001; Cook et al., 2002). This assumption may be violated, for instance, if a researcher fails to model existing non-linearity in the relation between the RV and the outcome. One way to address such an issue is to select smaller bandwidths, within which the specified linear model is more likely to be a close approximation, at the expense of lowering the power to detect small effect sizes (see Section 2.2.2). The third assumption is that no other factor exists that might cause the discontinuity at the cutoff point. This follows from the conceptualization of RD as local randomization. If baseline characteristics of participants also contributes to the discontinuities, then this assumption has been violated (D. S. Lee & Lemieux, 2010). In case of violation, one could enter those baseline characteristics as additional covariates into the baseline model and obtain adjusted LATE estimates.

Additionally, because conventional RD analysis is conducted using ordinary least squares (OLS) estimation, the following assumptions are also made. It is assumed that the outcomes is linearly related to the predictors, the error terms are normally distributed with a mean of zero and are homoscedastic, (i.e., the variance of the error terms is not dependent on any predictors), and observations are independent.

2.2.2 Bandwidth Selection for RD Models

Although RD analyses can be conducted with the full sample of data, researchers often prefer to select a smaller sub-sample such that the RV scores fall within a bandwidth around the pre-specified cutoff value. This is especially beneficial when the functional form that relates the outcome and RVs changes near the tails of the RV distribution. That is, even when the relation is globally nonlinear, a linear model can still be a good approximation in the neighborhood of the cutoff. Choosing a bandwidth involves a trade-off between 1) introducing bias from a bandwidth that is too wide due to the failure of the linear approximation and 2) losing precision from a bandwidth that is too narrow due to the limited sample size. Furthermore, the use of bandwidths also limits the generalizability of the treatment effect to only those with RV scores near the cutoff. There are many options for choosing the optimal bandwidth (e.g., Bloom, 2012; G. Imbens & Kalyanaraman, 2012) ranging from visual checks of graphs to empirical approaches.

A popular empirical approach is cross-validation (G. W. Imbens & Lemieux, 2008; D. S. Lee & Lemieux, 2010), which compares the ability of the estimator to predict outcomes for each sample observation at different bandwidth values. The bandwidth with the greatest prediction power is chosen for the final model. While it is difficult to know that a chosen bandwidth has reduced bias to an acceptable level, the more robust RD findings are in relation to different bandwidths, the more confidence can be placed in them.

Another option, which also incorporates a check of the modeled functional form (see Section 2.2.1) involves sequentially dropping different amounts of the outermost data and comparing the results to the original functional form (R. Jacob, Zhu, Somers, & Bloom, 2012). Finally, G. W. Imbens and Lemieux (2008) also recommend testing different cutoff values. Identifying other cutoff values that result in significant treatment effects would call into question the validity of the results obtained with the original cutoff point.

2.2.3 The Use of Covariates in RD Models

Including covariates in RD models has its root in viewing RD as an RCT in the vicinity of the cutoff point (Cattaneo et al., 2015; D. S. Lee, 2008), as the use of covariates is regularly employed in RCTs. Applied researchers often augment their RD models with covariates in order to increase the precision of the LATE estimator. Covariates may also reduce bias in the treatment effect estimator as they account for differences in observed characteristics between participants above and below the cutoff (Frölich, 2007). A covariate-adjusted LATE estimator is consistent with the standard RD treatment effect and results in more valid inference under minimal assumptions. In sharp RD designs, the only requirement is that covariates have equal conditional expectations above and below the cutoff point. If an interaction between the covariate and RV is present, an extra assumption is required for valid inference: The partial effect of the covariate on the potential outcome is homogenous, which is hard to justify in practice (Calonico, Cattaneo, Farrell, & Titiunik, 2016).

In education research, common covariates include demographics such as gender, age, and ethnicity, as well as assessment scores. For example, an investigation on the effect of financial aid offer on attending college (Van der Klaauw, 2002) included demographics, student grades and standardized test scores as covariates. Similarly, Calcagno and Long (2008) conducted their RD analysis both using covariates (i.e., demographics and a measure of language proficiency) and not using covariates. Their results were very similar across both methods, although, as expected, the standard errors were smaller when using the covariates.

2.2.4 Quantifying Generalizibility of Local Treatment Effect

In a conventional RD design the estimated LATE at the cutoff enjoys high internal validity; however, the external validity, or generalizability, of the estimate to values away

from the cutoff is limited. Bloom (2012) presents three views on the generalization of RD findings from the "strict-constructionist" view, which holds that without additional assumptions, the ATE is only identified for participants locally at the cutoff to the "more-expansive" view, which suggests measurement error in the RV facilitates the generalization of the LATE. Typifying this expansive view, Lee and Lemieux (2010) state the RD estimand can be interpreted as a " weighted [ATE] where the weights are directly proportional to the ex ante likelihood that an individual's realization of [the running variable] will be close to the threshold." They further state that the ex ante likelihood at the individual level is not obtainable from a single measure of the RV. Extra information such as "reliability", "a second test measurement", or "other covariates" that can predict the assignment is needed to generalize the LATE to a larger population.

Note that item-level data from which the RV scores are calculated are analogous to the "second test measurement". Therefore, it is possible to estimate the probability density/mass of the RV at the individual level, which consequently opens up the possibility of generalizing the LATE. Notably Angrist and Rokkanen (2015) proposed finding suitable exogenous covariates conditional on which the RV and the potential outcomes are mean-independent. Rokkanen (2015) established that the LV in the measurement model that governs the ORV satisfies this conditional independence assumption and, consequently, justified the generalizability of the LATE when a LV measurement model is explicitly and correctly specified. The proposed LV modeling approach to RD analysis builds on the "more-expansive" view of D. S. Lee and Lemieux (2010). Furthermore, it shares the same feature with Rokkanen (2015) in that the LVs underlying the RV satisfy the necessary conditional independence assumption of Angrist and Rokkanen (2015).

2.2.5 Disentangling the Heterogeneity of the Treatment Effect

One goal in many applications of RD analyses is to examine whether the LATE in the conventional RD model varies along additional variables. Harking back to D. S. Lee and Lemieux (2010), the LATE can be expressed as a weighted average of individual treatment effects across the entire population when the RV is not precisely measured. In other words, measurement error, or equivalently, the latent construct behind the ORV, can be a source of heterogeneity. However, D. S. Lee and Lemieux (2010)'s discussion was confined to a simplistic additive error model and thus, albeit insightful, is not directly applicable to structured measurement models, such as IRT models, in which the observed scores are typically nonlinearly mapped onto the LV. To unravel the heterogeneity in the LATE due to measurement error, it is crucial to fully specify a measurement model and empirically confirm its compatibility with the observed data using fit assessments for factor-analytic types of models, which have been extensively studied in the literature of psychometrics (e.g. Bock & Aitkin, 1981; Browne & Cudeck, 1993; Maydeu-Olivares & Joe, 2005, 2006; Rubin & Thayer, 1983).

2.2.6 Handling Measurement Error in Variables

Suppose that a study aims to evaluate a math tutoring program targeting at-risk high school students. A reasonable choice of the RV may be the math midterm score. However, because the midterm exam score is affected by the composition of test items, the score is not generalizable beyond the particular test administration and can only be treated as a noisy realization of true math proficiency. When the LATE needs to be estimated with respect to math proficiency — a latent construct — we essentially run into an error-in-variable problem.

It is known that measurement error under specific assumptions (e.g., linearity, additive,

symmetric, normal, independent) results in attenuated correlations and regression coefficients (e.g., Spearman, 1904). Furthermore, random measurement error, which does not follow these assumption, adds to the variability in the data and the size and magnitude of the bias in parameter estimates can vary. However, measurement error in the ORV has not been much discussed in the RD literature. This is likely because the bias in parameter estimates is not a problem when the interpretation of the LATE is confined to the scale of the ORV, instead of any latent construct underlying the ORV. Such interpretations could be of interest when the ORV is measured without error, e.g., exact age, or if it is not of high priority to generalize inferences about the treatment beyond the specific measurement of the construct being used. In the later case, the usefulness of the proposed model mainly resides in its convenience in probing the heterogeneity and generalizability of the LATE, as previously discussed. However, when inferences need to be made with respect to the latent construct, results from conventional RD analyses are not always valid.

The impact of measurement error on parameter estimates varies depending on which variable in the RD model is compromised as well as how large the measurement error is. First, it is known that measurement error in the outcome variable affects the standard errors of regression coefficients — and thus subsequent inferences such as hypothesis tests and confidence intervals. Second, measurement error in a predictor results in attenuated coefficient estimates for that predictor. Third, an error-contaminated covariate in the analysis of covariance (ANCOVA) incurs invalid statistical inferences for the regression coefficients of other (error-free) predictors (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). That is to say, the LATE parameter is at risk in an RD analysis even if it is only the covariates that are contaminated and the RV itself is free from measurement error.

Furthermore, as previously stated, measurement error can attenuate standardized effect sizes. A standardized effect size (e.g., Cohen's d) can be calculated by dividing the mean difference between two groups (e.g., treatment and control group) by the pooled standard

deviation of the groups. While measurement error will have no systematic effect on the mean difference between the groups, it will increase the variability in the scores and, consequently, increase the denominator in the standardized effect size calculation. This will result in an attenuated standardized effect size, which misrepresents the effect of the intervention being studied.

In addition to addressing issues related to measurement error, using LVs in RD analyses further allows researchers to benefit from modern measurement theory (e.g., IRT). Firstly, different forms of tests or instruments (e.g., Year 1 vs Year 2 exams) can be used to measure the latent construct underlying outcomes, running variables or covariates. In such cases, the LATE can be estimated when there are a subset of common items (i.e., anchor items) to link the different forms. Secondly, the LV framework allows measurement parameters estimated in previous studies to be reused in the current analysis, which increases the comparability of LATEs across studies, as the LATE can be interpreted with respect to the universal scale defined by the LVs. Finally, it is noted that factor-analytic techniques are typically used to explore/confirm the underlying structure of the construct in test development. The LV framework carries consistent definitions of the constructs to the subsequent RD analysis.

2.3 Latent Constructs as Predictors in RD Analyses

Many education interventions and programs target at-need groups. As such, they have existing eligibility criteria with a threshold for receiving the intervention, making RD designs a popular choice. Often these criteria are measures of latent constructs (e.g., math, ability, motivation).

Latent constructs are measured by collecting data on response variables such as continuous or categorical item responses. Responses on the items can be summed or averaged to create observed variable scores, or the relation between the responses and the latent con-

struct can be explicitly modeled to calculate LV scores (Thissen & Wainer, 2001) as in the IRT framework. In each case, the variable scores represent a participant's level of a construct of interest. These scores can then be used as predictors or outcomes in subsequent models (e.g., as the RV in an RD model). In education research, proxies for latent constructs are regularly used as RVs: School assessment scores (B. A. Jacob & Lefgren, 2004a; Chay et al., 2005; Chiang, 2009), entrance exam scores (Ding & Lehrer, 2007), standardized test scores (B. A. Jacob & Lefgren, 2004b; Van der Klaauw, 2008), course grades (Ou, 2010), and subject specific assessment scores (Goodman, 2008; Matsudaira, 2008).

Such RVs are especially popular in RD analyses of remedial education programs. Colleges will often give incoming students a "placement exam" to determine which level of a course is most appropriate for each student. These exams, which use standardized tests created by testing companies or individual universities, measure a latent construct (i.e., ability in a specified subject area). Thresholds for student placement are determined at the state or college level, and the values are generally known by the students before taking the exam. As such, students are sometimes able to negotiate placement in a course-level different than their assessment score indicated. Consequently, the majority of studies in this area are fuzzy RD designs which use the RD with instrumental variables approach to estimate treatment effects by using the expected discontinuity as an instrument for the actual discontinuity (Boatman, 2012; Boatman & Long, 2010; Calcagno & Long, 2008; Hodara, 2012; Martorell & McFarlin Jr, 2011; Scott, 2015). In this way, a treatment effect is estimated for compliers only, i.e., participants that received the treatment condition to which they were assigned.

Moreover, placement criteria are often based on more than just the placement exam score (e.g., faculty evaluation, other assessments, demographics). Melguizo, Bos, Ngo, Mills, and Prather (2016) investigated the effect of a remedial math program on student

achievement. In order to place students into a math course, a composite score was calculated for each student including score on the placement exam, scores on other assessments, student demographics, and letters of recommendation. The researchers found that fewer than five percent of the students were placed in a course different than the one their placement exam score alone would have indicated. As such, they performed a standard RD analysis based on the placement exam scores only and excluded the participants who were not in the course their placement exam score indicated.

2.3.1 Estimation of RD models with Latent Predictors

As mentioned in Chapter 1, RD studies using latent constructs as predictors use a multi-stage estimation approach, while only addressing the latter stage of the analyses (i.e., the RD model). Recall that in order to measure latent constructs, data on response variables are used to calculate either observed variable score or LV scores. In both cases, these scores are then used as the RV in an RD model. However, calculating LV scores with one model and using those scores in a second model may still result in misleading conclusions as the estimated LV scores are erroneously treated as known quantities in the second stage of estimation (Bollen, 1989; Skrondal & Laake, 2001). Consequently, estimating the relation among observed response variables, latent constructs, and outcomes should be conducted in one stage of estimation whenever feasible. While there has been a move in many fields towards a one-stage LV modeling framework (e.g., Bartholomew & Knott, 1999; Muthén, 2002; Rabe-Hesketh & Skrondal, 2004; Kaplan, Kim, & Kim, 2009), such a move has not been explored in the RD literature.

2.3.2 A Single-Level Latent RD Model

In this section, I first present a simple form of single-level RD analysis with a latent variable that is measured by the observed assignment variable, then illustrate how the generalizability and heterogeneity of the LATE can be examined and quantified with respect to observed RV scores by estimating the two regression lines with respect to the latent regressor.

A Single-level Latent Regression Discontinuity Model

When it is of interest to estimate the LATE with regards to a latent running variable $\theta_{r,i}$, the observed running variable X_i in the conventional RD model (Equations 2.3 and 2.4) is replaced by $\theta_{r,i}$, which yields:

$$Y_i = T_i f_1(\theta_{r,i}) + (1 - T_i) f_0(\theta_{r,i}) + \varepsilon_i$$

and

$$Y_i = \beta_0 + \beta_1 \theta_{r,i} + \beta_2 T_i + \beta_3 \theta_{r,i} T_i + \varepsilon_i,\ i = 1, \ldots, n. \tag{2.6}$$

While Y_i, ε_i, $f_1(\cdot)$, $f_0(\cdot)$, and the regression coefficients are defined in the same fashion as in Equation 2.4, the treatment indicator T_i, however, cannot be simply defined as $1\{\theta_{r,i} \geq c\}$, because $\theta_{r,i}$ is not observed and thus cannot be used for assigning participants into treatment versus control groups. Let $\hat{\theta}_{r,i}$ be an observed "proxy" of $\theta_{r,i}$, such as the summed score or a factor score. Note that $\hat{\theta}_{r,i}$ and $\theta_{r,i}$ may not be on the same scale. There is measurement error when the conditional distribution of $\hat{\theta}_{r,i}$ given $\theta_{r,i}$ is non-degenerate for some $\theta_{r,i}$. We can define $T_i = 1\{\hat{\theta}_{r,i} \geq c\}$ and proceed with treatment assignment. The LATE at $\theta_{r,i}$ is defined as $\text{LATE}(\theta_{r,i}) = \beta_2 + \beta_3 \theta_{r,i}$.

Suppose that the latent running variables $\theta_{r,i}$, $i = 1, \ldots, n$, are i.i.d. $N(0,1)$, each of

which is manifested by a collection of observed response variables $X_{ik}, k = 1, \ldots, K$. A measurement model characterizes the dependency of X_{ik} on $\theta_{r,i}$. For example, a dichotomous X_{ik}, i.e., $X_{ik} \in \{0, 1\}$, can be generated from a two-parameter logistic model (2PL; Birnbaum, 1968):

$$P\{X_{ik} = 1 | \theta_{r,i}; a_k, c_k\} = \frac{\exp(a_k \theta_{r,i} + c_k)}{1 + \exp(a_k \theta_{r,i} + c_k)}, \tag{2.7}$$

in which a_k and c_k denote the item slope and intercept parameters for item k, respectively. Similarly, the linear-normal factor analysis model can be used for continuous response variables, and the nominal response model can be used for nominal item responses.

An Illustrative Example

The linear, latent RD model (Equation 2.6) and its connection with the conventional RD model are illustrated in Figure 2.1. Binary item responses to 16 items were generated from a 2PL model. Summed scores are used as $\hat{\theta}_{r,i}$, and the treatment indicator is defined as $T_i = 1\{\hat{\theta}_{r,i} \geq 6\}$. The reliability coefficient alpha is approximately 0.85 under the data generating model, which is often regarded as "sufficiently high." While the conventional RD theory requires a continuous RV, in practice discrete RVs such as integer-valued test scores, as in this example, are often used.

By the tower property of conditional expectation, $E(Y_i^{[g]} | \hat{\theta}_{r,i}) = E[E(Y_i^{[g]} | \theta_{r,i}, \hat{\theta}_{r,i}) | \hat{\theta}_{r,i}], g = 0, 1$, which further equals to $E[E(Y_i^{[g]} | \theta_{r,i}) | \hat{\theta}_{r,i}]$ assuming conditional independence between potential outcomes, $Y_i^{[g]}$, and the ORV, $\hat{\theta}_{r,i}$, given the LRV, $\theta_{r,i}$. Consequently, the LATE in the conventional RD can be reinterpreted as

$$E(Y_i^{[1]} | \hat{\theta}_{r,i} = c) - E(Y_i^{[0]} | \hat{\theta}_{r,i} = c). \tag{2.8}$$

with the assumption that $P\{\hat{\theta}_{r,i} = c\} > 0$.

The conditional expectations in Equation 2.8 are taken with respect to the posterior distribution of the LRV, $\theta_{r,i}$, given the ORV, $\hat{\theta}_{r,i}$, which is governed by the measurement model parameters (e.g., a_k's and c_k's in the 2PL model); hence, the $LATE(\hat{\theta}_{r,i})$ is essentially $E(LATE(\theta_{r,i})|\hat{\theta}_{r,i} = c)$, a posterior average of the LATE with respect to the LRV $(\theta_{r,i})$. In the top panel of Figure 2.1, the quantities $E(Y_i^{[1]}|\hat{\theta}_{r,i})$ and $E(Y_i^{[0]}|\hat{\theta}_{r,i})$ are displayed as circles connected by solid lines. Note that those expected values computed from the latent RD model are not necessarily aligned with the fitted values from the conventional RD model (dashed lines), even at the cutoff value. The difference is traced to model misspecification as well as extrapolation on the left-hand side of the cutoff due to a discrete $\hat{\theta}_{r,i}$. The latent RD model reveals the heterogeneity intrinsic to the LATE at the cutoff $\hat{\theta}_{r,i} = c$ as illustrated in the second panel of Figure 2.1 where the shaded area represents the middle 90% of the posterior distribution of the LRV at the cutoff. A range of the LATE for this middle 90% of participants with an ORV at the cutoff $\hat{\theta}_{r,i} = 6$ can be calculated as

$$[LATE_{q.05}, LATE_{q.95}] = f_1(\theta_{r,i}) - f_0(\theta_{r,i}) : \theta_{r,i} \in [q_{.05}, q_{.95}], \tag{2.9}$$

where $q_{.05}$ and $q_{.95}$ are the $5th$ and $95th$ quantiles of the posterior distribution for individuals with an ORV at the cutoff. A range of the LATE at other quantiles of the LRV at the cutoff can also be computed.

Note, the RD design has perfect separation of treatment and control groups at the cutoff on the ORV (top panel of Figure 2.1) but imperfect separation on the LRV (bottom panel of Figure 2.1). The nonempty intersection between the two groups' LV ranges allows us to generalize the LATE to an interval of LV values. For example, the ATE within the ± 1 unit

interval around the cutoff $\hat{\theta}_{r,i} = c$, which we will refer to as $ATE_{c\pm1}$, can be calculated by,

$$ATE_{c\pm1} = E[f_1(\theta_{r,i}) - f_0(\theta_{r,i})|\hat{\theta}_{r,i}] \in [c-1, c+1] \tag{2.10}$$

As with Equation 2.8, it is assumed that $P\{\hat{\theta}_{r,i} \in [c-1, c+1]\} > 0$.

The imperfect separation between treatment and control observed in the bottom panel of Figure 2.1 also resembles a "fuzzy" RD design. This is a feature of the latent RD model. While the design is a sharp RD with respect to the ORV (top panel), it will always be a fuzzy RD with respect to the LRV. Furthermore, the separation in the bottom panel does not conform to any additive measurement error model and thus cannot be interpreted as a simple randomization in the neighborhood of the cutoff $\hat{\theta}_{r,i} = c$ (e.g., D. S. Lee & Lemieux, 2010). The validity of the generalized LATE is contingent upon the goodness-of-fit of the parametric latent RD model. As with conventional RD analysis, a simple parametric form (e.g., linear model) often approximates the data generating model better when the ORV values are restricted to a narrower bandwidth.

The LATE with respect to the ORV can also be computed at ORV values away from the cutoff (see the top panel of Figure 2.1) by replacing c with the desired ORV value in Equation 2.8. Furthermore, an ATE can be calculated with respect to a range of LRV values only. For example, the ATE among the bottom $p\%$ of participants in the population of the latent construct can be calculated as,

$$ATE_p = \frac{\int_{-\infty}^{z_p} \Big(f_1(\theta) - f_0(\theta) \Big) \phi(\theta) d\theta}{\int_{-\infty}^{z_p} \phi(\theta) d\theta} \tag{2.11}$$

where z_p is the pth quantile of the standard normal distribution (i.e., the distribution of the LRV).

Morell, Yang, and Liu (2019) explored the performance of the a single-level latent

RD model with a latent RV, latent covariate, and latent outcome. They found the model parameters were well recovered with short (5-item) and moderate (15-item) test lengths using sample sizes of 500 and 1000.

Additional Assumptions of the Latent Regression Discontinuity Model

Besides applying the conventional RD assumptions (see Section 2.2.1), the latent RD model (Equation 2.6) requires the following additional assumptions.

First, the measurement model is assumed to be correctly specified, rendering the LATE with respect to the latent variable meaningful. While the assumption is somewhat strong, its validity can be assessed by various model fit diagnostics developed for factor analysis and IRT models (e.g., Bentler, 1990; Browne & Cudeck, 1993; Joe & Maydeu-Olivares, 2010; Y. Liu, Yang, & Maydeu-Olivares, 2019a; Y. Liu & Maydeu-Olivares, 2014; Maydeu-Olivares & Liu, 2015; Maydeu-Olivares & Joe, 2005, 2006).

Second, the latent construct being measured is assumed to be the sole cause of the discontinuity.

Third, independence of potential outcomes $Y_i^{[g]}$, $g = 0, 1$, and the observed score $\hat{\theta}_{r,i}$ conditional on the latent assignment variable $\theta_{r,i}$ is assumed. In other words, the association between $Y_i^{[g]}$ and $\hat{\theta}_{r,i}$ is fully explained by $\theta_{r,i}$. Statistics based on residual cross-product moments, similar in spirit to Y. Liu and Maydeu-Olivares (2014), can be utilized to formally test conditional independence.

Fourth, normality and homoscedasticity of the error terms ε_i, $i = 1, \ldots, n$, is assumed because parameters in the latent RD model are typically estimated by ML. These distributional assumptions on the error terms can be examined via residual analysis analogous to that in mixed-effects regression.

2.4 Modeling Data Collected in Multilevel Settings

The sampling design often used in education settings is typically not simple random sampling in which independent samples are drawn from a single-level population of participants. Instead, cluster sampling, in which clusters (e.g., school districts, schools, classrooms, etc.) are first sampled and then participants (e.g., teachers, students) are randomly drawn within each sampled cluster, is more common. Studies of interventions or programs conducted in multilevel settings can be classified into two general types (Seltzer, 2004). In the block design, sampling occurs at the block (or cluster) level and individuals within the same block are assigned to treatment versus control groups (e.g., the effect of an intervention on student achievement is studied by sampling schools and assigning students within the schools to treatment or control). In the randomized cluster design, clusters themselves are assigned to different treatment conditions (e.g., all students in a school receive the intervention and all students in another school receive the control condition).

Data collected from such designs provide researchers with both benefits and challenges. While researchers can explore research questions that could not be answered with data collected via simple random sampling, standard modeling techniques cannot be used with clustered data due to assumption violations. Individuals within a cluster may be more similar to each other compared to individuals within a different cluster. Such data violate the assumption of independence of observations necessary for conducting common statistical analyses (e.g., ordinary least squares regression), which can result in underestimated standard errors of regression coefficients and, consequently, inflated Type-I error rates (Hox, Moerbeek, & Van de Schoot, 2017; Raudenbush & Bryk, 2002; Singer, Willett, Willett, et al., 2003). When cluster sizes are medium to large, even small dependencies in the data will result in large biases in the standard errors (Walsh, 1947). Historically, such data was analyzed by either aggregation (i.e., analyzing the data at the cluster level) or disaggregation

(i.e., ignoring the nested structure of the data) and using standard regression or analysis of variance methods. However, analyzing data from different levels (i.e., individual-level and cluster-level) at a common level results in both statistical and conceptual problems.

The statistical problems include a loss of power and inflation of the Type-I error rate. When data are aggregated, individual-level data are combined (e.g., averaged) to make fewer cluster-level units. Consequently, individual-level information is lost and the statistical analysis loses power. Conversely, when data are disaggregated, the information cluster-level data is amplified, and a few cluster-level units are treated as a large number of individual-level units. The true sample size for these variables is the number of cluster-level units; using the sample size at the disaggregated individual-level results in spurious significant results for standard statistical tests (Hox et al., 2017; Snijders & Bosker, 2012).

The conceptual problems occur when a researcher analyzes the data at one level but makes conclusions at a different level. The ecological fallacy occurs when aggregated data is interpreted at the individual level. For example, examining the relation between family income and school achievement and concluding students from wealthy families will perform better in school. The relation between variables estimated at the cluster-level will be stronger than the relation that exists at the individual level (Robinson, 1950). Similarly, analyzing disaggregated data and making inferences at the cluster-level results in the atomistic fallacy. For example, examining the relation between numbers of hours students spend on homework and student achievement and concluding if schools required more hours of homework, their students would have better academic outcomes (Hox et al., 2017).

In order to avoid these statistical and conceptual problems when analyzing clustered data, any dependency of observations must be accounted for using statistical methods. These can be grouped into two categories: design-based methods (Neyman, 1934), which use single-level models that adjust the standard errors of parameter estimates (e.g., regression coefficients) to accommodate violations of the assumption of independence of

observations, and model-based approaches (Fisher, 1955), which include the data clustering in the analytic model. Deciding between the two approaches is based on the research questions of interest (Bauer & Sterba, 2011; Stapleton, McNeish, & Yang, 2016).

Design-based approaches are appropriate when research questions do not address differences across clusters (e.g., "Does the intervention have an effect on student achievement?") and the researcher's concern is getting appropriate estimates of the standard errors of the model parameters. Such methods include generalized estimating equations (Liang & Zeger, 1986), linearization (Binder & Roberts, 2003), and resampling methods such as jackknife and bootstrapping (Efron, 1960). These methods have been shown to adequately adjust standard errors under a variety of sampling conditions (Heeringa, West, & Berglund, 2017; McNeish, 2014; McNeish, Stapleton, & Silverman, 2017). Furthermore, they have the added benefit of being simpler to implement and interpret than model-based approaches.

The model-based approach, i.e., fitting a multilevel model, also called a "hierarchical linear model" (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), a "mixed-effects" or "mixed linear model" (Littell, Henry, & Ammerman, 1998), and a "random coefficient model" (Kreft & De Leeuw, 1998), is appropriate when researchers are interested in the relation between cluster level characteristics and individual level outcomes (e.g., "Does the effect an intervention has on student achievement differ by clusters?"). Multilevel regression allows for unobserved heterogeneity in the outcome to be modeled via random intercepts and unobserved heterogeneity in the effects of predictors on the outcome to be modeled as random slopes (Rabe-Hesketh & Skrondal, 2004).

The current study focuses on multilevel modeling approaches to handling clustered data to the exclusion of design-based approaches as it aims to provide a tool for researchers with complex research questions that require multilevel RD models. The rest of this section describes the multilevel modeling framework, assumptions of multilevel models, and

important modeling decisions.

The multilevel regression model assumes the use of a hierarchical data set with individuals, i, nested in clusters, j, an outcome variable measured at the individual level, and predictors included at both levels:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_2 X_{2ij} + \varepsilon_{ij} \tag{2.12}$$

In this model, the intercept, β_{0j} and the regression slope β_{1j} each have a j subscript, as it is assumed these coefficients vary across clusters. These coefficients are referred to as the random intercept and a random slope, respectively. In contrast, β_2, does not have a subscript j and is not assumed to vary across clusters. The regression coefficients can be fixed or allowed to vary depending on the hypothesized model. Note ε_{ij} is the error term at the individual level and has a variance of σ^2. Equation 2.12 is referred to as a "level 1 equation." The corresponding level 2 equations are

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}Z_j + u_{1j} \\ \beta_2 &= \gamma_{20} \end{aligned} \tag{2.13}$$

The u_{0j} and u_{1j} are random residual error terms at the cluster level. More random effects, u_j terms, can be added to the model based on the research question. The variance of u_{0j} and u_{1j} are specified as τ_0^2 and τ_1^2, respectively; the covariance between the terms is specified as τ_{01}. Note that the γ terms do not have j subscripts and are not assumed to vary across clusters; they are referred to as fixed coefficients. Finally, Z_j is a predictor at the cluster level and therefore does not have an i subscript. The level 1 (Equation 2.12) and level 2

(Equation 2.13) models can also be expressed as a combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} Z_j X_{1ij} + u_{1j} X_{1ij} + \gamma_{20} X_{2ij} + \varepsilon_{ij}. \tag{2.14}$$

Note the cross-level interaction between the level-1 predictor X_{1ij} and the level-2 predictor Z_j. This is an additional benefit, and complication, of multilevel modeling. Cross-level interactions allow researchers to examine if a relation at a lower level, e.g., the relation between student motivation and student outcomes, depends on a predictor at a higher level, e.g., classroom size; however, the interpretation of a cross-level interaction is complicated by centering decisions, which is discussed in more detail below (see Section 2.4.2). The combined form of the model also illustrates how the variance in the outcomes is decomposed as variance between (τ_0) and variance within (σ^2)clusters, which results in unbiased estimates of the standard errors of regression coefficients (Snijders & Bosker, 2012; Raudenbush & Bryk, 2002; Hox et al., 2017). We can further calculate the ratio of variance between clusters to the total variance, which is called the *intraclass correlation* (ICC).

$$\rho = \frac{\tau_0}{\tau_0 + \sigma^2} \tag{2.15}$$

The ICC estimates the proportion of the total variance explained by the clustering structure in the population; it can also be interpreted as the expected correlation between two randomly drawn units from the same cluster (Hox et al., 2017). In practice, this value is calculated based on an intercept-only model, rendering each variance term the unexplained variance at each level. After adding predictors to the model, the proportion of variance explained, R^2, can be calculated as in single-level regression; however, as there are multiple levels, an R^2 is calculated at each level.

The proportion of variance explained at level-1 is (Raudenbush & Bryk, 2002):

$$R_1^2 = \frac{\sigma_b^2 - \sigma_m^2}{\sigma_b^2} \tag{2.16}$$

where σ_b^2 is the residual variance at the individual level for the baseline model (intercept-only model) and σ_m^2 is the individual level residual variance for the comparison model. Similarly, the The proportion of variance explained at level-2 is (Raudenbush & Bryk, 2002):

$$R_2^2 = \frac{\tau_{0|b} - \tau_{0|m}}{\tau_{0|b}} \tag{2.17}$$

where $\tau_{0|b}$ is the residual variance at the cluster level for the baseline model and $\tau_{0|m}$ is the cluster level residual variance for the comparison model.

2.4.1 Assumptions of Multilevel Models

The assumptions of multilevel models are as follows. The residual errors at level 1, ε_{ij} are assumed to be independent and normally distributed with mean 0 and variance σ^2 for every level-1 unit i within each cluster, j, i.e., the variance of the level-1 residuals is assumed to be the same in each cluster. To test the assumption, residuals at level 1 should be checked for normality and for a constant variance across level-1 predictors. The predictors at level 1 are independent of ε_{ij} and level-2 predictors are independent of u_j. Note that independence between omitted predictors and the residual error terms cannot be tested, although, residuals can be plotted against variables not included in the model to examine if there is a relation between the two. The random error terms at level 2, u_j have a multivariate normal distribution with mean 0 and a constant covariance matrix. The random error vectors are independent among level-2 units. This assumption can be by checking the distribution of the level-2 residuals. The error terms at level 2 are independent of the error

terms at level 1, and the predictors at each level are uncorrelated with the random effects at the other levels.

2.4.2 Centering Predictors

A key decision for researchers using multilevel modeling is how to enter predictors into the model, e.g., uncentered, centered at the mean, centered at a meaningful value. Centering a predictor is subtracting a constant, often the mean, from every value of a variable. This scaling changes the interpretation of the intercept in both single-level and multilevel models. For example, if every predictor is centered at its mean, the intercept is interpreted as the expected value of the outcome when all predictors have their mean values. This is particularly useful when a value of zero for a predictor is not meaningful. In multilevel regression, centering also changes the interpretation of the variances of the intercept and slopes. Furthermore, in addition to being uncentered or centered at a meaningful value, predictors in multilevel regression can be centered at their overall mean, i.e., grand mean centering, or at their cluster-level means, i.e., group mean centering.

While group mean centering has been recommended by Raudenbush and Bryk (2002) and Enders and Tofighi (2007), under certain conditions, and entering predictors uncentered or centered at a meaningful value can be useful depending on the scale of the predictors and the research question (Hox et al., 2017), grand mean centering is generally the most popular and highly recommended approach (Bickel, 2007; Enders & Tofighi, 2007; Hox et al., 2017; Kelley, Evans, Lowman, & Lykes, 2017; Kenny, Kashy, Bolger, et al., 1998; Paccagnella, 2006). Grand mean centering allows for ease of interpretation as the intercept and variance components are interpreted as the expected values when all predictors equal their means. Furthermore, a model with all predictors mean-centered is equivalent to the uncentered model. In contrast, a model with group mean centered predictors is

not equivalent to its uncentered (or grand mean centered) counterpart (Enders & Tofighi, 2007; Paccagnella, 2006). This is because instead of subtracting a single value from all raw scores, different values are subtracted depending on the cluster, which consequently removes information about between-cluster differences from the model.

As previously stated, the centering decision also has implications for interpreting cross-level interactions. Modeling interactions in multilevel analysis is not as straightforward as in single-level analysis. When there is a significant interaction effect in the model, the effect of the coefficient for the interaction and the coefficients for the direct effects of the predictors in the interaction must be interpreted as a system (Aiken, West, & Reno, 1991; Jaccard, Wan, & Turrisi, 1990). That is, the regression coefficients of the direct effects have different meaning in the presence and absence of a significant interaction effect. When the model includes an interaction, the regression coefficient of each of the predictors in the interaction is interpreted as the expected value of that regression slope when the other predictor is equal to zero (Hox et al., 2017).

Centering decisions for latent predictors can also affect model parameter estimates. Entering latent predictors at level-1 as uncentered or using grand mean centering can result in attenuated interaction effects at level 1, while group mean centering can attenuate level-2 interaction effects (Ryu, 2012). In contrast, group mean centering at level 1 and grand mean centering at level 2 is preferred (Asparouhov & Muthén, 2019).

2.4.3 RD Designs in Multilevel Settings

RD in multilevel settings, generally referred to as "multilevel RD," can also described using Seltzer (2004) taxonomy. The RD counterpart to the randomized cluster design is the hierarchical RD (HRD; see Rhoads & Dye, 2016) design, in which the RV is measured at a higher level and the outcome variable is measured at a lower level. An example would be a

Table 2.1: Multilevel RCT and RD Designs

Design		Design Features
Cluster-Level Assignment	RCT	Level 2 units are randomly assigned to treatment or control
	RD	Running variable is measured at level 2 and outcome is measured at level 1
Individual-Level Assignment	RCT	Level 1 units (within each level 2 unit) are randomly assigned to treatment or control
	RD	Running variable and outcome are measured at level 1

study in which schools are eligible for an intervention based on the percentage of students with a below grade-level reading score, and the success of the intervention is measured by student achievement. The treatment assignment happens at the school level, while the outcomes are measured at the student level. The block design in RD is called a multisite RD (MRD; see Rhoads & Dye, 2016) design. In the MRD design, the RV must be measured at the same level as the outcome, but participants exist in clusters which are treated as random effects. For example, a study in which students district-wide are eligible for a tutoring program based on their scores on an assessment, and the effect of the program is measured by student achievement, is a two-level MRD. These designs are summarized in Table 2.1. Furthermore, multiple RVs can be used (e.g., Wong, Steiner, & Cook, 2013), the RVs can exist at both levels (e.g., school eligibility and student eligibility), and the cutoff can vary by cluster; this work focuses on the single RV scenarios in which either school or student level cutoff is used to assign treatment.

The conventional HRD and MRD models using observed variables are presented below.

The Observed Variable HRD Model

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}(X_{r,ij} - X_{r,j}) + \beta_{2j}(X_{c,ij} - X_{c,j}) + \varepsilon_{ij}, \tag{2.18}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}T_j + u_{0j}, \tag{2.19}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}T_j,$$

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

The equivalent combined equation is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{r,ij} - X_{r,j}) + \gamma_{01}T_j + \gamma_{11}(X_{r,ij} - X_{r,j})T_j + \gamma_{20}(X_{c,ij} - X_{c,j}) + u_{0j} + \varepsilon_{ij}, \tag{2.20}$$

in which Y_{ij} is the observed outcome variable, $X_{r,ij}$ is the ORV and $X_{c,ij}$ is the observed covariate, $X_{r,j}$ and $X_{c,j}$ are the group means for the ORV and the observed covariate. γ_{00} is the mean for the control clusters, γ_{10} is the relation between the ORV and the outcome variable, γ_{01} is the LATE, γ_{11} is the interaction between the ORV and the treatment, γ_{20} is the relation between the observed covariate and the outcome, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and $u_{0j} \sim N(0, \tau_0^2)$.

The Observed Variable MRD Model

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}(X_{r,ij} - X_{r,j}) + \beta_{2j}T_{ij} + \beta_{3j}(X_{r,ij} - X_{r,j})T_{ij} + \beta_{4j}(X_{c,ij} - X_{c,j}) + \varepsilon_{ij} \tag{2.21}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + u_{0j} \tag{2.22}$$

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

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\beta_{3j} = \gamma_{30}$$

$$\beta_{4j} = \gamma_{40}$$

The equivalent combined equation is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{r,ij} - X_{r,j}) + \gamma_{20}T_{ij} + \gamma_{30}(X_{r,ij} - X_{r,j})T_{ij} + \gamma_{40}(X_{c,ij} - X_{c,j}) + u_{0j} + u_{2j}T_{ij} + \varepsilon_{ij}, \tag{2.23}$$

in which Y_{ij} is the observed outcome variable, $X_{r,ij}$ is the ORV and $X_{c,ij}$ is the observed covariate, $X_{r,j}$ and $X_{c,j}$ are the group means for the ORV and the observed covariate. γ_{00} is the mean for the control clusters, γ_{10} is the relation between the ORV and the outcome variable, γ_{20} is the LATE, γ_{30} is the interaction between the ORV and the treatment, γ_{40} is the relation between the observed covariate and the outcome, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and and $(u_{0j}, u_{2j})^T \sim N(0, \mathbf{T})$ where

$$\mathbf{T} = \begin{pmatrix} \tau_0^2 & \tau_{20} \\ \tau_{20} & \tau_2^2 \end{pmatrix}.$$

2.5 Regression Discontinuity Model Estimation

2.5.1 Multi-Stage Estimation

The studies presented in this chapter use a two stage estimation approach where scores for the latent constructs are obtained either by observed variable methods, such as summing or taking the average of item responses, or by latent variable methods, such as obtaining *expected a posteriori* scores, *maximum a posteriori* scores, or ML scores from an IRT model. The scores obtained from any of these methods are then used in the RD model, which is estimated using ordinary least squares regression. However, using observed variable methods to calculate assignment variable scores does not account for the measurement error in the scores, which can attenuate model parameter estimates (Spearman, 1904). While calculating LV scores is a more appropriate method, the direct use of such LV scores in subsequent statistical analyses can result in biased estimates of model parameters and standard errors (Hoijtink & Boomsma, 1995; Lu, Thomas, & Zumbo, 2005; Mislevy, Johnson, & Muraki, 1992). This occurs because the sampling variability in the estimated LVs is not carried over into the structural model. Instead, the LVs are treated as known values, which results in an underestimation of standard errors and biased parameter estimates. Popular methods to address this limitation in using LV scores include the use of plausible values (Mislevy, 1984, 1985; Mislevy & Sheehan, 1983; Yang & Seltzer, 2015) and adjusting the standard errors of parameter estimates (Y. Liu, Yang, & Maydeu-Olivares, 2019b; Wang, Weiss, & Su, 2019). However, as previously discussed, a one-stage estimation framework in which the relation between the latent construct and the observed item responses is estimated at the same time as the RD treatment effect is the preferred approach.

2.5.2 Single-Stage Estimation

One class of single-stage estimation approaches, fully Bayesian estimation, draws samples that target the posterior distribution via Markov chain Monte Carlo (MCMC) procedures. Such approaches to the RD design have been widely developed in the econometrics methodological literature (D. S. Lee & Card, 2008; Rau, 2011; Branson, Rischard, Bornn, & Miratrix, 2019; Karabatsos & Walker, 2015; Geneletti, Ricciardi, OâĂŹKeeffe, & Baio, 2019). Fully Bayesian estimation has been used in estimating the effect of school attendance (Chib & Jacobi, 2016; Li, Mattei, & Mealli, 2015b), employment (Lalive, 2007), and access to health care (Venkataramani, Bor, & Jena, 2016; Vandenbroucke & Le Cessie, 2014; Cawley & Talbot, 2006). While a Bayesian estimation approach can be used with the proposed models, there are two limitations that make its implementation less desirable. First, in the proposed models response patterns are used to estimate measurement model parameters. When there are item responses that are not observed in the data, Bayesian estimation will be inappropriate. Second, the use of MCMC estimation requires many decisions (e.g., selection of prior distributions, determination of "burn-in" period, convergence diagnostic of chains). As education researchers tend to be less familiar and less comfortable with Bayesian estimation approaches, their use in estimating the proposed models makes the adoption of the proposed models more cumbersome.

Single-stage frequentist estimation approaches can be categorized broadly into limited-information and full-information methods. When used with categorical data, as in the current study, limited-information methods can be less burdensome to estimate. This is because these methods utilize a contingency table instead of individual item response, so that only the univariate and bivariate margins are used. However, this computational ease is at the cost of omitting the higher-order associations that exist among item responses, making full-information estimation the preferred method (Maydeu-Olivares & Joe, 2005). Full-

information estimation methods utilize all available response data, which has the added advantage of being a model-based approach to handling missing data (see e.g., Enders, 2010).

However, the use of full-information estimation in multilevel LV modeling with categorical response data presents computational difficulties that can quickly make such estimation of the model infeasible. Popular full-information estimation schemes in LV modeling, such as marginal ML with the expectation maximization (EM; Bock & Aitkin, 1981) algorithm, rely on high-dimensional numerical integration of the LVs by use of quadrature points. This makes them subject to the "curse of dimensionality" (Bellman, 1957) whereby the number of quadrature points needed increases exponentially as a function of the number of LVs in the model. For example, a model with 5 LVs, where each dimension is integrated by 11 quadrature points, results in $11^5 = 161,051$ quadrature points.

As the limitation of implementing these algorithms in high-dimensional LV models is the number of quadrature points needed, alternative algorithms were developed so that fewer nodes would be needed for estimation. One such alternative, adaptive quadrature rules (Q. Liu & Pierce, 1994; Naylor & Smith, 1982), strategically places quadrature nodes under areas of the LVs' posterior distribution with higher density. This allows fewer quadrature nodes to achieve comparable estimates to fixed quadrature point approaches. While adaptive quadrature rules may allow for a larger number of LVs than marginal ML with EM methods, it is still limited by the number LVs that can be estimated in a model. A second alternative algorithm scheme is the Monte Carlo expectation maximization algorithm (Meng & Schilling, 1996), which uses sampling to obtain quadrature points. However, as parameter estimates near the ML, Monte Carlo EM requires an increasingly large number of random draws (Cai, 2010a, 2010b; Gu & Kong, 1998), resulting in a similar limitation to marginal ML estimation via the EM algorithm.

The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm was developed as an

alternative to these algorithms, and it has been implemented in a context similar to the current work to estimate contextual effects through nonlinear multilevel latent variable models (Yang & Cai, 2014). In the next section, I describe the general principles of this algorithm.

2.5.3 The MH-RM Algorithm

Let f be the general symbol for probability density/mass functions and ξ be all the parameters in the measurement and structural models. The ML estimator of ξ can be obtained by maximizing the log-likelihood function $\log f(\mathbf{X}, \mathbf{Y}|\xi)$, where $\mathbf{X}$ is the observed "predictor-side" data and $\mathbf{Y}$ is the observed "outcome-side" data. Because of the high-dimensional integrations involved in the likelihood function, the MH-RM algorithm is used, which is a stochastic counterpart of the Newton-Raphson type algorithm. The MH-RM algorithm approximates the observed data likelihood by first using Fisher's Identity for data augmentation via a Metropolis-Hastings (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) sampler and then employs the Robbins-Monro (RM; Robbins & Monro, 1951) stochastic approximation algorithm. Because the MH-RM algorithm does not use numerical integration, it allows for a large number of latent variables to be modeled simultaneously. As such, the proposed model can still be estimated with little added complexity in the case of multidimensional latent variable predictors or the inclusion of several additional latent covariates.

Each iteration, $t = 1, \ldots, T$, of the MH-RM algorithm consists of three steps: Stochastic Imputation, Stochastic Approximation, and Robbins-Monro Update.

Step 1: Stochastic Imputation. Let $\mathbf{Z}$ collect all random effects, latent RVs (LRVs) and latent covariates, which form the "missing data". These random effects and latent variables are drawn from a Markov chain that targets the posterior predictive distribution of

missing data given the observed data $(\mathbf{X},\mathbf{Y})$, i.e., $f(\mathbf{Z}|\mathbf{X},\mathbf{Y},\xi^{(t)})$, where $\xi^{(t)}$ is the current estimates of model parameters at iteration $t : t = 1,\ldots,T$. At each iteration, M_t sets of complete data are formed as follows:

$$\{\mathbf{Y},\mathbf{X},\mathbf{Z}_m^{(t+1)}; m = 1,\ldots,M_t\} \tag{2.24}$$

Step 2: Stochastic Approximation. The gradient vector of the complete data log-likelihood function is defined as:

$$\mathbf{s}(\xi|\mathbf{Y},\mathbf{X},\mathbf{Z}) = \frac{\partial}{\partial\xi}\log f(\mathbf{Y},\mathbf{X},\mathbf{Z}|\xi). \tag{2.25}$$

By Fisher's (1925) identity, the gradient of the observed data log-likelihood is the expectation of $\mathbf{s}(\xi|\mathbf{X},\mathbf{Y},\mathbf{Z})$ over the posterior distribution $f(\mathbf{Z}|\mathbf{X},\mathbf{Y},\xi)$, that is,

$$\frac{\partial}{\partial\xi}\log f(\mathbf{X},\mathbf{Y}|\xi) = \int \mathbf{s}(\xi|\mathbf{X},\mathbf{Y},\mathbf{Z}) f(\mathbf{Z}|\mathbf{X},\mathbf{Y},\xi) d\mathbf{Z}. \tag{2.26}$$

Equation 2.26 evaluated at $\xi^{(t)}$ gives the direction of steepest ascent. For LV models, direct evaluations of the observed data gradient is computationally expensive, if not completely infeasible. Nevertheless, given a set of imputed missing data $\mathbf{Z}_m^{(t+1)}$, $m = 1,\ldots,M_t$, Equation 2.26 suggests the following Monte Carlo estimates of the observed data gradient:

$$\tilde{\mathbf{s}}_{t+1} = \frac{1}{m_t}\sum_{m=1}^{M_t} \mathbf{s}(\xi^{(t)}|\mathbf{Y},\mathbf{X},\mathbf{Z}_m^{(t+1)}). \tag{2.27}$$

$\tilde{\mathbf{s}}_{t+1}$ gives a noise-corrupted version of $s(\xi^{(t)}|\mathbf{X},\mathbf{Y})$.

Step 3: Robbins-Monro Update. A recursive approximation of the conditional expectation of the complete-data information matrix at the $(t+1)th$ iteration (e.g., Cai, 2008; Gu

& Kong, 1998) is computed as:

$$\Gamma_{t+1} = \Gamma_t + g_t \left[\frac{1}{M_t} \sum_{m=1}^{M_t} H(\xi^{(t)} | \mathbf{X}, \mathbf{Y}, \mathbf{Z}_m^{(t+1)}) - \Gamma_t \right], \tag{2.28}$$

where the complete data information matrix is

$$\mathbf{H}(\xi | \mathbf{X}, \mathbf{Y}, \mathbf{Z}) = -\frac{\partial^2}{\partial \xi \partial \xi'} \log f(\xi | \mathbf{X}, \mathbf{Y}, \mathbf{Z}). \tag{2.29}$$

Parameters are updated recursively:

$$\xi^{(t+1)} = \xi^{(t)} + g_t \Gamma_{t+1}^{-1} \tilde{\mathbf{s}}_{t+1}, \tag{2.30}$$

in which g_t;$t > 1$ is a decaying sequence of gain constants, which can be defined to filter out noise across iterations. In practice, different gain constants are used in different stages of the MH-RM algorithm. A single gain constant value is used across Stage I iterations in order to stabilize parameter estimates around the ML estimates. Similarly, Stage II uses a single gain constant value in all iterations with the goal of obtaining good starting values for the next stage. Stage III has gain constants defined by $\sum_{t=1}^{\infty} g_t = \infty$ and $\sum_{t=1}^{\infty} g_t^2 < \infty$. The algorithm is terminated once the minimum change in a parameter estimate is below a desired threshold for a window of iterations (Cai, 2008).

Standard Error Estimation. Standard errors of structural and measurement parameters can be estimated by the Louis formula (Louis, 1982), which can be applied either iteratively or after convergence. The former produces estimates of standard errors as a by-

product of the MH-RM iterations, whereas the later requires post-convergence MH sampling steps. When applied iteratively, the gradient vector is approximated recursively as:

$$\hat{\mathbf{s}}_{t+1} = \hat{\mathbf{s}}_t + g_t\{\tilde{\mathbf{s}}_{t+1} - \hat{\mathbf{s}}_t\}, \tag{2.31}$$

where $\tilde{\mathbf{s}}_{t+1}$ is defined as

$$\tilde{\mathbf{s}}_{t+1} = \frac{1}{M_t}\sum_{m=1}^{M_t} \mathbf{s}(\xi^{(t)}|\mathbf{Y},\mathbf{X},\mathbf{Z_m}^{(t+1)}) \tag{2.32}$$

in each iteration of the MH-RM algorithm. A Monte Carlo estimate of the conditional expectation is defined as follows:

$$\tilde{\mathbf{G_t}} = \frac{1}{M_t}\sum_{m=1}^{M_t}\left[\mathbf{H}(\xi^t|\mathbf{X},\mathbf{Y},\mathbf{Z}_m^{t+1}) - \mathbf{s}(\xi^t|\mathbf{X},\mathbf{Y},\mathbf{Z}_m^{t+1})\left[\mathbf{s}(\xi^t|\mathbf{X},\mathbf{Y},\mathbf{Z}_m^{t+1})\right]'\right] \tag{2.33}$$

which is further stabilized by recursive approximation,

$$\hat{\mathbf{G}}_{t+1} = \hat{\mathbf{G}}_t + g_t\{\tilde{\mathbf{G}}_{t+1} - \hat{\mathbf{G}}_t\} \tag{2.34}$$

The observed data information is then approximated as

$$I_{t+1} = \hat{\mathbf{G}}_{t+1} + \hat{\mathbf{s}}_{t+1}\hat{\mathbf{s}}_{t+1}^{'} \tag{2.35}$$

2.6 Summary

The RD design is a popular choice for education researchers as it allows for the estimation of causal effects without the need for random assignment. When participants are assigned to treatment groups based on where their scores on a running variable (RV) fall in relation to an exogenous cutoff value, a local average treatment effect (LATE) can be calculated. While the LATE enjoys high internal validity, it has limited generalizability, as it is only applicable to participants with RV scores near the cutoff. This limits the usefulness of the model as the programs or interventions that researchers are interested in evaluating treat individuals with a wide range RV values. Furthermore, education research regularly measures latent constructs, which cannot be directly observed. Instead, data are collected on a set of indicators, e.g., survey item responses. While psychometricians use latent variable (LV) models to account for the relation between item responses and the latent construct, conventional RD analysis treats these item responses as observed variables by calculating sums or averages of the responses. This approach has several disadvantages: Inferences made based on observed scores are only applicable to that data collection and are not generalizable to other administrations of that survey or instrument; Results from multiple studies using different instruments to examine the same latent construct cannot be combined; The LATE can only be estimated with respect to the observed RV and not to the underling latent construct. As such, using LVs in place of observed variables would be preferred. The use of observed RVs in conventional RD analysis is due to the belief that measurement error in the RV works to convert the RD design to a local randomized control trial and allows researchers to examine heterogeneity and generalizability of the LATE (D. S. Lee & Lemieux, 2010). However, the simplistic additive error model discussed by D. S. Lee and Lemieux (2010) does not apply to modern measurement models. Therefore, in order to be able to explore these features of the LATE, a measurement model must be

specified for the RV. When LVs are used in a structural model, it is preferable to estimate the relation between item responses and the latent construct at the same time as the structural regression parameters (e.g., the LATE). When item responses are categorical, standard full-information estimation approaches become infeasible as the number of latent variables in the model increases. Due to the large number of latent variables in the proposed models, I use the MH-RM algorithm, which avoids numerical integration.

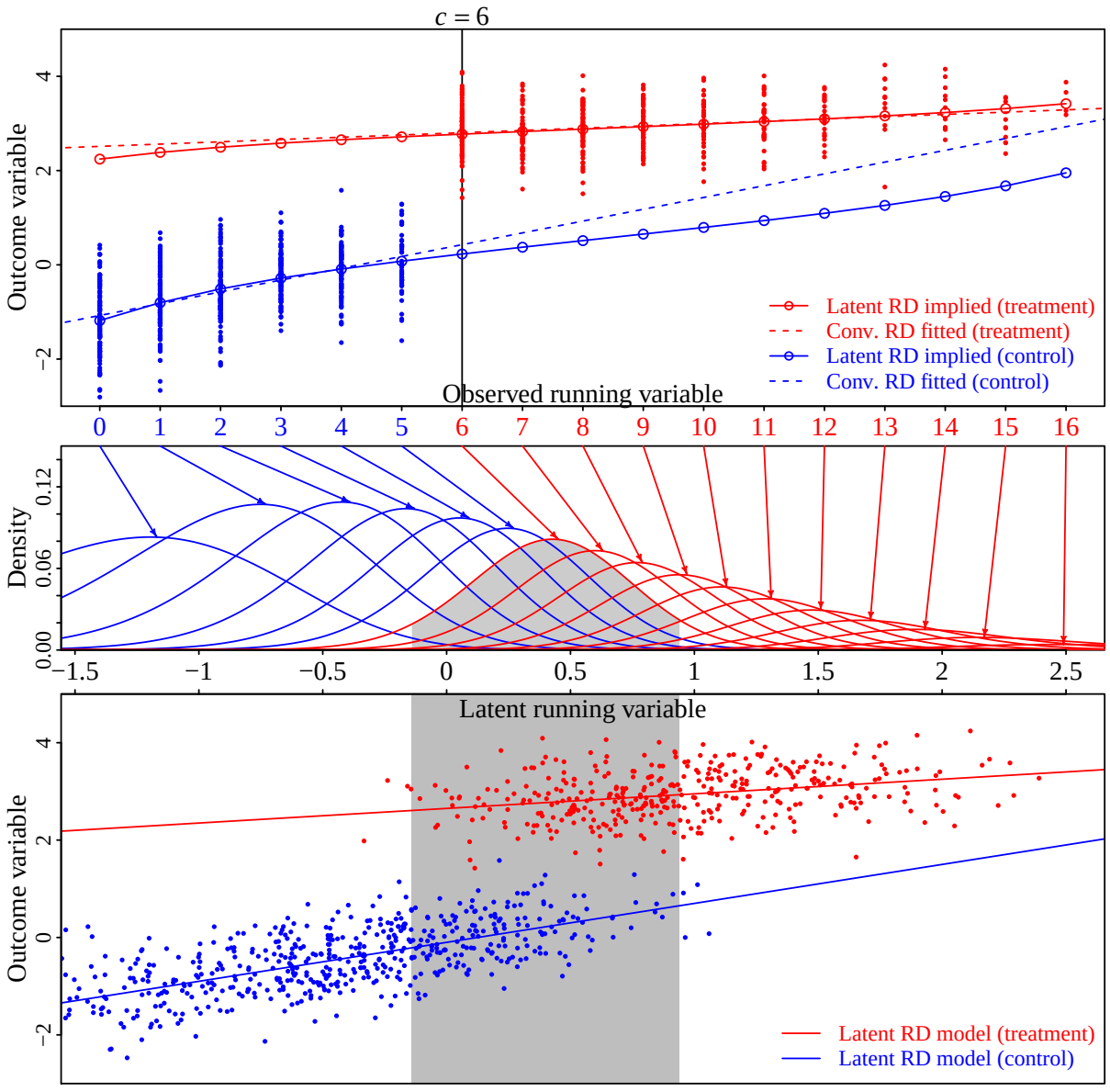

Figure 2.1: Conventional and latent regression discontinuity model. Top panel: Outcome variable values plotted against observed running variable scores. Treatment and control groups are shown in red and blue colors, respectively. Group-specific expected outcomes at each observed running variable score (i.e., summed score) were obtained from a fitted conventional RD model (dashed) and the true latent RD model (solid). Middle panel: Density functions of the latent running variable at different summed-score levels. The shaded area represents the middle 90% latent running variable values conditional on a summed score of 6, the cutoff value. Bottom panel: Outcome variable values plotted against the true latent running variable values.

Chapter 3: Multilevel RD Models with Latent Variables and Estimation

This chapter first illustrates the proposed models for multilevel LV RD modeling, then describes the use of the MH-RM algorithm to estimate the proposed models.

3.1 A General Multilevel RD Model with Latent Variables

In this section, the general measurement and structural model with notations is first presented and then two major exemplary latent RD models are described.

3.1.1 Measurement Model

Let i index individuals and j index clusters: Individuals are nested within clusters, such that $j = 1,\ldots,J$ and $i = 1,\ldots,I_j$. For each individual i within cluster j, let θ_{ij} be a vector of latent running variables and covariates, measured by a collection of K indicator variables $\mathbf{X}_{ij} = (X_{ijk} : k = 1,\ldots,K)$. In particular, it is assumed that

$$\theta_{ij} = \theta_j + \delta_{ij}, \tag{3.1}$$

in which θ_j denotes the latent cluster mean (level-2 LVs) and δ_{ij} denotes the individual deviation around the cluster means (level-1 LVs). Pooling across all i's within cluster j, let

$\mathbf{X}_j = (\mathbf{X}_{ij} : i = 1, \ldots, I_j)$. A "predictor-side" measurement model specifies the likelihood of $\mathbf{X}_j$ conditional on θ_j:

$$f_j(\mathbf{x}_j|\theta_j) = \prod_{i=1}^{I_j} \prod_{k=1}^{K} f_k(x_{ijk}|\theta_{ij}) \tag{3.2}$$

in which $\mathbf{x}_j = (x_{ijk} : i = 1, \ldots, I_j; k = 1, \ldots, K)$ denotes a realization of $\mathbf{X}_j$. The double product in Equation 3.2 results from the assumption of independence among $X_{ij1}, \ldots, X_{ijK}$ conditional on θ_{ij}—a key assumption of factor analytic models (McDonald, 1981). By default, both the level-1 and level-2 components of θ_{ij} are assumed to follow multivariate normal distributions.

The conditional likelihood of x_k given θ_{ij}, i.e., $f_k(x_{ijk}|\theta_{ij})$, can take various functional forms depending on the data type. The current study focuses on dichotomous data as the use of cognitive assessments, in which item responses can either be correct or incorrect, is popular in education research. The most commonly used IRT models for such data are the 2PL model and the three-parameter logistic (3PL) model (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Lord & Novick, 1968). These models differ in the number of properties that are estimated for each item. The 3PL model requires a larger sample size to properly estimate three parameters for every item (De Ayala, 2013). Furthermore, several studies have pointed out issues regarding the estimation and interpretation of the 3PL's "third parameter", the guessing parameter (Lord, 1974, 1975, 1980; Hambleton et al., 1991; Holland, 1990). To avoid introducing these additional issues, the current work focuses on multilevel counterpart of the 2PL model (Equation 2.7) described in Section 2.3.2, which allows each item to have a unique difficulty value and discrimination value.

$$f_k(x_{ijk}|\theta_{ij}) = \frac{\exp[x_{ijk}(\mathbf{a}_k^\top(\theta_j + \delta_{ij}) + c_k)]}{1 + \exp(\mathbf{a}_k^\top(\theta_j + \delta_{ij}) + c_k)}, \tag{3.3}$$

in which $\mathbf{a}_k$ denotes the slopes, which are assumed to be invariant across level 1 and level 2, and c_k are the intercepts. The θ_j is distributed as $N \sim (\mathbf{0}, Var(\theta_j))$ and δ_{ij} follows a standard normal distribution.

Let η_{ij} be the latent outcome variable, $\mathbf{Y}_{ij} = (Y_{ijl} : l = 1,\ldots,L)$ be the associated indicator variables, $\eta_j = (\eta_{ij} : i = 1,\ldots,I_j)$, and $\mathbf{Y}_j = (\mathbf{Y}_{ij} : i = 1,\ldots,I_j)$. An "outcome-side" measurement model is governed by the following likelihood of $\mathbf{Y}_j$ conditional on η_j

$$f_j(\mathbf{y}_j) = \prod_{i=1}^{I_j} \prod_{l=1}^{L} f_l(y_{ijl}|\eta_{ij}), \tag{3.4}$$

in which $\mathbf{y}_j = (y_{ijl} : i = 1,\ldots,I_j; l = 1,\ldots,L)$ denotes an instance of $\mathbf{Y}$, and $f_l(y_{ijl}|\eta_{ij})$ denotes the conditional likelihood of Y_{ijl} given η_{ij}. As with the predictor-side measurement model, the 2PL model is considered for categorical response data.

3.1.2 Structural Model

$$\eta_{ij} = T_{ij} f_1(\theta_{ij}, \mathbf{u}_j) + (1 - T_{ij}) f_0(\theta_{ij}, \mathbf{u}_j) + \varepsilon_{ij}, \tag{3.5}$$

in which $\mathbf{u}_j$ denotes the random effects. The treatment indicator often takes the form $T_{ij} = 1\{\mathbf{X}_j \in B\}$, in which B is a pre-specified region in the space of $\mathbf{X}_j$. As the assignment may happen at the cluster level, we allow T_{ij} to be dependent on X-side indicators $\mathbf{X}_{ij}$ for all $i = 1,\ldots,I_j$. ε_{ij}'s are i.i.d. $N(0,\sigma^2)$ error terms, assumed to be independent to θ_{ij}'s and $\mathbf{X}_{ij}$'s. Two instances of Equation 3.5 are discussed in Sections 3.1.2 to 3.1.2. Note that the LVs in the models are automatically grand mean centered.

Hierarchical Regression Discontinuity Models

In a HRD design, the treatment assignment occurs at the cluster level; therefore, the treatment indicator is denoted T_j with the subscript i dropped from the notation in the general RD model (Equation 3.5). Recall, the LV θ_{ij} collects the latent running variables $\theta_{r,ij} = \theta_{r,j} + \delta_{r,ij}$ and the latent covariates $\theta_{c,ij} = \theta_{c,j} + \delta_{c,ij}$. The LATE is a fixed effect in the HRD design; random effects can be specified for the intercept and the coefficients for the level-1 components of the predictor-side latent variables. Substituting the observed variables in Equations 2.18 and 2.20, the structural part of the random-intercept-only HRD model can be expressed as

$$\begin{aligned}
\text{Level 1: } \eta_{ij} &= \beta_{0j} + \beta_1(\theta_{r,ij} - \theta_{r,j}) + \beta_2(\theta_{c,ij} - \theta_{c,j}) + \varepsilon_{ij}, \qquad (3.6)\\
\text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}T_j + u_{0j},\\
\beta_{1j} &= \gamma_{10} + \gamma_{11}T_j,\\
\beta_{2j} &= \gamma_{20}
\end{aligned}$$

The equivalent combined equation is

$$\eta_{ij} = \gamma_{00} + \gamma_{10}(\theta_{r,ij} - \theta_{r,j}) + \gamma_{01}T_j + \gamma_{11}(\theta_{r,ij} - \theta_{r,j})T_j + \gamma_{20}(\theta_{c,ij} - \theta_{c,j}) + u_{0j} + \varepsilon_{ij}, \quad (3.7)$$

in which η_{ij} is the latent outcome variable, $\theta_{r,ij}$ and $\theta_{c,ij}$ are the latent running variable and covariate, which, per convention in IRT, each have normal distributions, $\theta_{r,j}$ and $\theta_{c,j}$ are the cluster means, which each have standard normal distributions, γ_{00} is the fixed effect intercept, and the other terms are defined as in the observed HRD model. For the

purpose of identification, σ^2 is fixed at 1, to prevent Heywood cases. Furthermore, the factor standardization approach to model identification was chosen so that factor loadings could be estimated for each item. As such, γ_{00} is fixed at 0. The LATE($\theta_{r,ij}$) is calculated as $\gamma_{01} + \gamma_{11}\theta_{r,ij}$.

For simplicity of notation, all fixed effect coefficients not associated with a random term will be labeled as β as shown below:

$$\eta_{ij} = \gamma_{00} + \beta_1(\theta_{r,ij} - \theta_{r,j}) + \beta_2 T_j + \beta_3(\theta_{r,ij} - \theta_{r,j})T_j + \beta_4(\theta_{c,ij} - \theta_{c,j}) + u_{0j} + \varepsilon_{ij}, \quad (3.8)$$

A list of all measurement and structural parameters for the HRD model appears below.

- Measurement parameters:
  - $\theta_{r,ij} \sim N(0, 1 + Var(\theta_{r,j}))$: Latent, continuous RV for individual i in cluster j
  - $\theta_{c,ij} \sim N(0, 1 + Var(\theta_{c,j}))$: Latent, continuous covariate for individual i in cluster j
  - $\eta_{ij} \sim N(0, \sigma^2 + \tau_0^2)$: Latent, continuous outcome variable for individual i in cluster j
  - a_k: Slope for predictor item k
  - c_k: Intercept for predictor item k
  - a_l: Slope for outcome item l
  - c_l: Intercept for outcome item l
- Structural parameters:
  - γ_{00}: Regression intercept, fixed at 0 for model identification

- β_1: Relation between latent running variable and outcome variable
- T_j: Observed, dichotomous variable which records the treatment condition assigned at the cluster level based on the average cluster ORV
- β_2: Main effect of the treatment (LATE)
- β_3: Interaction between the latent running variable and the treatment effect
- β_4: Relation between latent covariate and outcome variable
- $\varepsilon_{ij} \sim N(0, \sigma^2)$: Individual-level residual
- σ^2: Individual-level residual variance, fixed at 1 for model identification
- $u_{0j} \sim N(0, \tau_0^2)$: Cluster-level residual
- τ_0^2: Cluster-level residual variance

Multisite Regression Discontinuity Models

Recall that a MRD design assigns individuals to treatment versus control groups within each cluster. It is, therefore, of interest to study not only the overall LATE but also the extent to which the LATE varies from cluster to cluster. As a result, at least two random effects are included in the baseline MRD model—one for the intercept, and the other for the LATE. Because the assignment happens within each cluster, the latent treatment indicator T_{ij} is a function of $\mathbf{X}_{ij}$. Substituting the observed variables in Equations 2.21 and 2.23, the structural part of the baseline MRD model is then

$$\text{Level 1: } \eta_{ij} = \beta_{0j} + \beta_{1j}(\theta_{r,ij} - \theta_{r,j}) + \beta_{2j}T_{ij} + \beta_{3j}(\theta_{r,ij} - \theta_{r,j})T_{ij} + \beta_{4j}(\theta_{c,ij} - \theta_{c,j}) + \varepsilon_{ij} \tag{3.9}$$

$$\begin{aligned}
\text{Level 2: } \beta_{0j} &= \gamma_{00} + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + u_{2j} \\
\beta_{3j} &= \gamma_{30} \\
\beta_{4j} &= \gamma_{40}
\end{aligned}$$

The equivalent combined equation is

$$\eta_{ij} = \gamma_{00} + \gamma_{10}(\theta_{r,ij} - \theta_{r,j}) + \gamma_{20}T_{ij} + \gamma_{30}(\theta_{r,ij} - \theta_{r,j})T_{ij} + \gamma_{40}(\theta_{c,ij} - \theta_{c,j}) + u_{0j} + u_{2j}T_{ij} + \varepsilon_{ij}, \tag{3.10}$$

in which η_{ij} is the latent outcome variable, $\theta_{r,ij}$ and $\theta_{c,ij}$ are the latent running variable and covariate, which, per convention in IRT, each have normal distributions, $\theta_{r,j}$ and $\theta_{c,j}$ are the cluster means, which each have standard normal distributions γ_{00} is the fixed-effect intercept, and the remaining coefficients are defined as in the observed MRD model.

As in the HRD model, $\varepsilon_{ij} \sim N(0, \sigma^2)$ with σ^2 fixed at 1 and γ_0 fixed at 0 for model identification. The fixed-effect part of the LATE (with respect to the LRV) is given by $\gamma_{20} + \gamma_{30}\theta_{r,ij}$). Cluster-specific LATEs can be obtained by combining the fixed-effect part with the empirical Bayes estimates of u_{0j} and u_{2j}. In cases when the number of clusters J is large (e.g., at least 55 with an ICC of 0.25; see Rhoads & Dye, 2016 for more details)

more random effects can be added to the baseline model. While in the HRD model, the RV must be measured at the individual level, in the MRD model, the RV may be measured at the individual level and aggregated to the cluster level, e.g., cluster assignment is based on average student reading ability, or at the cluster level itself, i.e., assignment of clusters is based on a measure of principal quality. The proposed MRD model may be applied under both conditions as the measurement model may be specified at either the individual or the cluster level.

For simplicity of notation, all fixed effect coefficients not associated with a random term will be labeled as β as shown below:

$$\eta_{ij} = \gamma_{00} + \beta_1(\theta_{r,ij} - \theta_{r,j}) + \gamma_{20}T_{ij} + \beta_3(\theta_{r,ij} - \theta_{r,j})T_{ij} + \beta_4(\theta_{c,ij} - \theta_{c,j}) + u_{0j} + u_{2j}T_{ij} + \varepsilon_{ij}, \tag{3.11}$$

A list of all measurement and structural parameters for the MRD model appears below.

- Measurement parameters:
  - $\theta_{r,ij} \sim N(0, 1 + Var(\theta_{r,j}))$: Latent, continuous RV for individual i in cluster j
  - $\theta_{c,ij} \sim N(0, 1 + Var(\theta_{c,j}))$: Latent, continuous covariate for individual i in cluster j
  - $\eta_{ij} \sim N(0, \sigma^2 + \tau_0^2 + 2\tau_{02}T_{ij} + \tau_2^2 T_{ij}^2)$: Latent, continuous outcome variable for individual i in cluster j
  - a_k: Slope for predictor item k
  - c_k: Intercept for predictor item k
  - a_l: Slope for outcome item l
  - c_l: Intercept for outcome item l

- Structural parameters:
  - γ_{00}: Regression intercept, fixed at 0 for model identification
  - β_1: Relation between latent running variable and outcome variable
  - T_{ij}: Observed, dichotomous variable which records the treatment condition assigned at the individual level based on the ORV
  - γ_{20}: Main effect of the treatment (LATE)
  - β_3: Interaction between the latent running variable and the treatment effect
  - β_4: Relation between latent covariate and outcome variable
  - $\varepsilon_{ij} \sim N(0, \sigma^2)$: Individual-level residual
  - σ^2: Individual-level residual variance, fixed at 1 for model identification
  - $(u_{0j}, u_{2j})^T \sim N(0, \mathbf{T})$ where

$$\mathbf{T} = \begin{pmatrix} \tau_0^2 & \tau_{02} \\ \tau_{02} & \tau_2^2 \end{pmatrix}.$$

    Cluster-level residuals
  - τ_0^2: Cluster-level residual variance
  - τ_2^2: Variance in cluster-level LATE
  - τ_{02}: Covariance between regression intercept and main effect of the treatment

Note, in the current work, the cutoff is not allowed to vary by cluster. While this is a plausible condition seen in applications where the cutoff rule is based on a field or federal standard, there are also applications in which the cutoff may vary by cluster, e.g., when each cluster has its own assignment rule. The proposed model may be applied under this

latter scenario; however, the interpretation of the LATE is complicated by the varying cutoff values. When there is a single cutoff value across clusters, an overall LATE is calculated. When cutoff values are allowed to vary, the overall LATE becomes an average treatment effect (ATE) as the neighborhood for which a treatment effect is being calculated varies across clusters. The interpretation of the treatment effect changes in that the effect is for participants with ORV values near the cutoffs at each cluster. An ATE that is generalizable to ORV values away from the cutoffs may still be calculated. Furthermore, while heterogeneity in the ATE may still be calculated, allowing cutoff values to vary across clusters introduces another source of heterogeneity in the ATE.

3.1.3 Generalizability and Heterogeneity of Treatment Effect

As discussed in Section 2.3.2, the LATE in the conventional RD model can be interpreted as

$$E(Y_i^{[1]}|\hat{\theta}_{r,i} = c) - E(Y_i^{[0]}|\hat{\theta}_{r,i} = c),$$

rendering the LATE with respect to the ORV, a posterior average of the LATE with respect to the LRV: $E(LATE(\theta_{r,i})|\hat{\theta}_{r,i} = c)$.

Quantifying the generalizability of the LATE and calculating the heterogeneity in the LATE due to measurement error in the ORV requires calculating the likelihood for a given summed score $\dot{x}$. For any IRT model with possible binary responses and items indexed by k, the likelihood of a given summed score, $\dot{x} = \sum_k x_{ik}$, is calculated by summing the likelihoods of all the individual response patterns that have that summed score:

$$f(\dot{x}, \theta) = \sum_{\sum_k x_{ik} = \dot{x}} f(\mathbf{x}_i, \theta). \tag{3.12}$$

As the number of items in a test increases, the number of response patterns also increases

and a "brute force" calculation of this likelihood becomes impossible. This is especially a concern for the HRD model as assignment occurs using cluster-level summed scores, which further increases the total number of response patterns used in the calculation. As an alternative to the brute force method, Lord and Wingersky (1984) described a recursive algorithm for the computation of summed score likelihoods for items with binary responses.

The summed score likelihood is computed as follows:

$$f_{[k]}(\dot{x},\theta) = f_k(0|\theta) \times f_{[k-1]}(\dot{x},\theta) + f_k(1|\theta) \times f_{[k-1]}(\dot{x}-1,\theta), \tag{3.13}$$

where the first two terms are the $\dot{x}$ endorsements in $k-1$ items and the last two terms are the $\dot{x}-1$ endorsements in $k-1$ items. Equation 3.13 calculates the summed score likelihood for the first k items, which can then be used to quantify the generalizability of the LATE and calculate the heterogeneity in the LATE due to measurement error in the ORV.

3.1.4 Observed and Complete Data Likelihoods

Recall from Section 2.5.3 that ξ collects all measurement and structural parameters in the models. The conditional density of $\mathbf{y}_{ij}$ is written as:

$$f_\xi(\mathbf{y}_{ij}|\eta_{ij}) = f_\xi(\mathbf{y}_{ij}|\theta_j, \delta_{ij}, \mathbf{u}_j, \varepsilon_{ij}) \tag{3.14}$$

where $\theta_j = (\theta_{r,j}, \theta_{c,j})$, $\delta_{ij} = (\delta_{r,ij}, \delta_{c,ij})$, and $\mathbf{u}_j$ contains only u_{0j} under the HRD model (see Equation 3.8) and contains (u_{0j}, u_{2j}) under the MRD model (see Equation 3.11). As previously stated, $\varepsilon_{ij} \sim N(0,\sigma^2)$ with σ^2 is fixed at 1. Integrating out ε_{ij} yields

$$f_{\xi_{ij}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\theta_j, \delta_{ij}, \mathbf{u}_j) = \int f_\xi(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\theta_j, \delta_{ij}, \mathbf{u}_j, \varepsilon_{ij}) f_\xi(\varepsilon_{ij}) d\varepsilon_{ij} \tag{3.15}$$

And integrating out δ_{ij} gives

$$f_\xi(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\theta_j, \mathbf{u}_j) = \int f_\xi(\mathbf{x}_{ij}|\delta_{ij}) f_\xi(\mathbf{y}_{ij}|\delta_{ij}, \theta_j, \mathbf{u}_j) f_\xi(\delta_{ij}) d\delta_{ij} \tag{3.16}$$

The conditional joint densities of $\mathbf{y}_j$ and $\mathbf{x}_j$ for group j is obtained by multiplying the conditional join densities for $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ in the same group:

$$f_\xi(\mathbf{y}_j, \mathbf{x}_j|\theta_j, \mathbf{u}_j) = \prod_{i=1}^{I_j} f_\xi(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\theta_j, \mathbf{u}_j) \tag{3.17}$$

Integrating out the level-2 latent variables, θ_j, and random coefficients, $\mathbf{u}_j$, results in

$$f_\xi(\mathbf{y}_j, \mathbf{x}_j) = \int \prod_{i=1}^{I_j} f_{xxi}(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\theta_j, \mathbf{u}_j) f_\xi(\theta_j) f_\xi(\mathbf{u}_j) d\theta_j d\mathbf{u}_j \tag{3.18}$$

Thus, the marginal distribution from which parameters can be estimated is obtained by integrating out all latent variables and random coefficients. Treating $\eta_{ij}, \theta_j, \delta_{ij}, \mathbf{u}_j$, and ε_{ij} as missing data, the complete data likelihood for J groups, where each group j contains I_j individuals is:

$$\prod_{j=1}^{J} \left\{ \prod_{i=1}^{I_j} \{ \prod_{l=1}^{L} f_\xi(\mathbf{y}_{ijl}|\delta_{ij}, \theta_j, \mathbf{u}_j, \varepsilon_{ij}) \} \{ \prod_{k=1}^{K} f_\xi(\mathbf{x}_{ijk}|\delta_{ij}) \} f_\xi(\delta_{ij}) f_\xi(\varepsilon_{ij}) \right\} \times f_\xi(\mathbf{u}_j) f_\xi(\theta_j), \tag{3.19}$$

where K are the number of item response variables used to measure each θ_{ij} and L are the

number of item response variables used to measure each η_{ij}.

3.2 Model Parameter Estimation

In the proposed study, full-information maximum likelihood (FIML) estimation of model parameters will be carried out in R version 3.4.4 (Team et al., 2018). Parameters of interest include the measurement parameters (e.g., item discrimination, item difficulty) and structural parameters (e.g., treatment effect). The next section describes how the MH-RM algorithm is implemented to obtain maximum likelihood estimates for the HRD and MRD models in the multilevel latent variable modeling framework. The implementation is the same under both proposed models except where otherwise noted.

3.2.1 MH-RM Algorithm Implementation

The Metropolis-Hastings Sampler

Recall from Section 2.5.3 that $\mathbf{X}$ is the observed predictor-side data, $\mathbf{Y}$ is the observed outcome-side data, and $\mathbf{Z}$ collects all random effects, latent running variables and latent covariates, which form the "missing data". As previously discussed, the first step of the MH-RM algorithm consists of the stochastic imputation of latent variables and random effects. This imputation process aims at sampling from the distribution of missing data given observed data, $F_{\xi}(\mathbf{Z}|\mathbf{X},\mathbf{Y})$, which is proportional to the complete data likelihood $L(\xi|\mathbf{Z},\mathbf{X},\mathbf{Y})$. It works as follows: 1) candidate values are generated using a random walk sampler, 2) the acceptance probabilities are evaluated, and 3) the candidate values are accepted or rejected. Recall from Equations 3.1, 3.8, and 3.11 the latent variables are θ_j, δ_{ij},

and η_{ij}, the random variables are ε_{ij} and $\mathbf{u}_j$, where $\mathbf{u}_j$ contains u_{0j} under the HRD model (see Equation 3.8) and contains (u_{0j}, u_{2j}) under the MRD model (see Equation 3.11). The parameters which are considered fixed in the population, i.e., all parameters except for latent variables and random effects, can be denoted ξ_f.

For model identification purposes, γ_{00} is fixed at 0 and the variance of ε_{ij} is fixed at 1, which allows all item factor loadings to be estimated. Per standard convention in IRT, $\theta_{,ij}$ and $\theta_{c,ij}$ are each defined with a mean of zero and variance 1. The random variables of interest are the individual-level latent predictors δ_{ij}, the group level latent predictors θ_j, the level-1 residuals, ε_{ij}, and the level-2 residuals $\mathbf{u}_j$, where $\mathbf{u}_j = u_{0j}$ in the HRD model and $\mathbf{u}_j = (u_{0j}, u_{2j})$ in the MRD model. As η_{ij} is calculated once the other parameter estimates are known, it is excluded. These random variables are treated as missing data and correspond to $\mathbf{Z}$, first introduced in Section 2.5.3. The level-1 random variables are collected in $\mathbf{Z}_{ij} = (\delta_{ij}, \varepsilon_{ij})$ and the level-2 random variables are collected in $\mathbf{Z}_j = (\theta_j, \mathbf{u}_j)$, and both are treated as missing data. The MCMC imputation procedure is constructed using the Metropolis-Hastings-within-Gibbs sampler developed by Yang and Cai (2014). Let $\mathbf{Z}_{ij}^t$ be the value of $\mathbf{Z}_{ij}$ in the tth iteration of the sampler with the following steps:

$$
\begin{aligned}
&\text{Draw}\mathbf{Z}_{1j}^t \sim f_{\xi_f}(\mathbf{Z}_{1j}|\mathbf{Z}_{2j}^{t-1}, \ldots, \mathbf{Z}_{Ij}^{t-1}, \mathbf{Z}_j, \mathbf{Y}, \mathbf{X}) \\
&\text{Draw}\mathbf{Z}_{2j}^t \sim f_{\xi_f}(\mathbf{Z}_{2j}|\mathbf{Z}_{1j}^t, \mathbf{Z}_{3j}^{t-1}, \ldots, \mathbf{Z}_{Ij}^{t-1}, \mathbf{Z}_j, \mathbf{Y}, \mathbf{X}) \\
&\vdots \\
&\text{Draw}\mathbf{Z}_{ij}^t \sim f_{\xi_f}(\mathbf{Z}_{ij}|\mathbf{Z}_{1j}^t, \ldots, \mathbf{Z}_{i-1j}^t \mathbf{Z}_{i+1j}^{t-1}, \mathbf{Z}_{Ij}^{t-1}, \mathbf{Z}_j, \mathbf{Y}, \mathbf{X}) \\
&\vdots \\
&\text{Draw}\mathbf{Z}_{Ij}^t \sim f_{\xi_f}(\mathbf{Z}_{Ij}|\mathbf{Z}_{1j}^t, \ldots, \mathbf{Z}_{I-1j}^t, \mathbf{Z}_j, \mathbf{Y}, \mathbf{X})
\end{aligned}
\tag{3.20}
$$

The acceptance probability of moving from state $\mathbf{Z}_{ij}$ to $\mathbf{Z}^*_{ij}$, is given by parameters ξ_f, observed data $\mathbf{Y}$ and $\mathbf{X}$, and missing data at level-2 $\mathbf{Z}_j$ is

$$\alpha(\mathbf{Z}_{ij},\mathbf{Z}^*_{ij}|\xi_f,\mathbf{Y},\mathbf{X},\mathbf{Z}_j)=min\left\{\frac{f_{\xi_f}(\mathbf{Y},\mathbf{X}|\mathbf{Z}^*_{ij})h_{1j}(\mathbf{Z}^*_{ij}|\mu_{1j},\Sigma_{1j})h_2(\mathbf{Z}_j|\mu_2,\Sigma_2)q(\mathbf{Z}^*_{ij},\mathbf{Z}_{ij})}{f_{\xi_f}(\mathbf{Y},\mathbf{X}|\mathbf{Z}_{ij})h_{1j}(\mathbf{Z}_{ij}|\mu_{1j},\Sigma_{1j})h_2(\mathbf{Z}_j|\mu_2,\Sigma_2)q(\mathbf{Z}_{ij},\mathbf{Z}_{ij})},1\right\} \tag{3.21}$$

where μ_{1j} and Σ_{1j} and μ_2 and Σ_j define the distribution of the level-1 and level-2 missing data, respectively (see Section 4.1.3), $q(\mathbf{Z}_{ij},\mathbf{Z}^*_{ij})$ is a transition density, and $\mathbf{Z}^*_{ij}$ is $\mathbf{Z}_{ij}+e_{ij}$, where e_{ij} follows a scaled multivariate standard normal distribution with number of dimensions equal to the number of latent variables, e.g., for the increment to $\mathbf{Z}_{ij}=(\delta_{ij},\varepsilon_{ij})'$ a set of e_{ij} is drawn from a scaled standard multivariate normal distribution $N_3(\mathbf{0},w^2\mathbf{I_3})$. The acceptance rate of the MH chain is tuned by adjusting the value of w.

Equation 3.21 can be reduced due to the symmetry of the increment density $q(\mathbf{Z}_{ij},\mathbf{Z}^*_{ij})=q(\mathbf{Z}^*_{ij},\mathbf{Z}_{ij})$ and because the density function related to level-2 missing data $h_2(\mathbf{Z}_j|\mu_j,\Sigma_s)$ is the same for current and candidate draws. This yields the form:

$$\alpha(\mathbf{Z}_{ij},\mathbf{Z}^*_{ij}|\xi_f,\mathbf{Y},\mathbf{X},\mathbf{Z}_j)=min\left\{\frac{f_{\xi_f}(\mathbf{Y},\mathbf{X}|\mathbf{Z}^*_{ij},\mathbf{Z}_j)h_1(\mathbf{Z}^*_{ij}|\mu_{1j},\Sigma_{1j})}{f_{\xi_f}((\mathbf{Y},\mathbf{X}|\mathbf{Z}_{ij})h_{1j}(\mathbf{Z}_{ij}|\mu_{1j},\Sigma_{1j})},1\right\} \tag{3.22}$$

The acceptance probabilities are calculated by Equation 3.22, and the candidates are accepted or rejected based on the evaluation.

Once the level-1 candidate draws are accepted or rejected, level-2 random effect candidates are generated. Let $\mathbf{Z}^t_j$ be the value of $\mathbf{Z}_j$ in the tth iteration of the sampler with the following steps:

$$
\begin{aligned}
&\text{Draw}\mathbf{Z}_1^t \sim f_{\xi_f}(\mathbf{Z}_1|\mathbf{Z}_2^{t-1},\ldots,\mathbf{Z}_J^{t-1},\{\mathbf{Z}_{ij}\}_{i=1}^{I_1},\mathbf{Y},\mathbf{X})\\
&\text{Draw}\mathbf{Z}_2^t \sim f_{\xi_f}(\mathbf{Z}_2|\mathbf{Z}_1^t,\mathbf{Z}_3^{t-1},\ldots,\mathbf{Z}_{/j}^{t-1},\{\mathbf{Z}_{ij}\}_{i=1}^{I_2},\mathbf{Y},\mathbf{X})\\
&\vdots\\
&\text{Draw}\mathbf{Z}_j^t \sim f_{\xi_f}(\mathbf{Z}_j|\mathbf{Z}_1^t,\ldots,\mathbf{Z}_{j-1}^t\mathbf{Z}_{j+1}^{t-1},\mathbf{Z}_J,\{\mathbf{Z}_{ij}\}_{i=1}^{I_j},\mathbf{Y},\mathbf{X})\\
&\vdots\\
&\text{Draw}\mathbf{Z}_J^t \sim f_{\xi_f}(\mathbf{Z}_J|\mathbf{Z}_1^t,\ldots,\mathbf{Z}_{J-1}^t,\{\mathbf{Z}_{ij}\}_{i=1}^{I_j},\mathbf{Y},\mathbf{X})
\end{aligned}
\tag{3.23}
$$

Level-2 random effects candidates are generated as $\mathbf{Z}_j^* = \mathbf{Z}_j + e_j$. As in level-1, e_j is drawn from a scaled standard multivariate normal distribution. Under the HRD model the distribution is $N_3(\mathbf{0}, w^2\mathbf{I_3})$, and under the MRD model the distribution is $N_4(\mathbf{0}, w^2\mathbf{I_4})$, corresponding to the number of level-2 random effects in each model. Now the likelihoods are evaluated at level-2 as $\mathbf{Z}_j$ are level-2 random variables. The acceptance probability of moving from $\mathbf{Z}_j$ to $\mathbf{Z}_j^*$ is calculated by:

$$
\alpha(\mathbf{Z}_j,\mathbf{Z}_j^*|_{,\mathbf{f}},\mathbf{Y},\mathbf{X},\mathbf{Z}_{ij}) = min\left\{\frac{\prod_{i=1}^{I_j} f_{,\mathbf{f}}(\mathbf{Y},\mathbf{X}|\mathbf{Z}_{ij}^*,\mathbf{Z}_j^*)h_2(\mathbf{Z}_j^*|\mu_2,\Sigma_j)}{\prod_{i=1}^{I_j} f_{,\mathbf{f}}(\mathbf{Y},\mathbf{X}|\mathbf{Z}_{ij},\mathbf{Z}_j)h_2(\mathbf{Z}_j|\mu_2,\Sigma_j)},1\right\} \tag{3.24}
$$

Alternating sampling the level-1 missing data conditional on level-2 missing data and then sampling level-2 missing data conditional on the level-1 missing data, the sequence of drawings in the MH sampler converge in distribution to $F_{\xi_f}(\mathbf{Z}|\mathbf{Y},\mathbf{X})$ (Gelfand & Smith, 1990; Geman & Geman, 1984).

Standard Error Estimation

Standard errors of structural and measurement parameters were shown to be underestimated when using the recursive approximation approach in a pilot study. Consequently, in the current study standard errors are estimated by the Louis formula (Louis, 1982), which is applied after convergence of fixed parameter estimates:

$$I_{\mathbf{X},\mathbf{Y}} = E_{\xi_f}\{\mathbf{H}(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)|\mathbf{X},\mathbf{Y},\mathbf{Z}\} \tag{3.25}$$

$$- E_{\xi_f}\{\mathbf{s}(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)\mathbf{s}^\top(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)|\mathbf{X},\mathbf{Y},\mathbf{Z}\},$$

$$+ E_{\xi_f}\{\mathbf{s}(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)|\mathbf{X},\mathbf{Y},\mathbf{Z}\}E_{\xi_f}\{\mathbf{s}^\top(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)|\mathbf{X},\mathbf{Y},\mathbf{Z}\}, \tag{3.26}$$

The first term in Equation 3.26 is calculated as

$$E_{\xi_f}\{\mathbf{H}(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)|\mathbf{X},\mathbf{Y},\mathbf{Z}\} \approx \frac{1}{v}\sum_{i=1}^{v}\mathbf{H}(\hat{\xi}_f,\mathbf{X}_i,\mathbf{Y}_i,\mathbf{Z}_i|\mathbf{X}_i,\mathbf{Y}_i,\mathbf{Z}_i), \tag{3.27}$$

and the second term in Equation 3.26 is calculated as

$$E_{\xi_f}\{\mathbf{s}(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)\mathbf{s}^\top(\mathbf{X},\mathbf{Y},\mathbf{Z},\xi_f)\mathbf{X},\mathbf{Y},\mathbf{Z}\} \tag{3.28}$$

$$\approx \frac{1}{v}\sum_{i=1}^{v}[\mathbf{s}(\hat{\xi}_f,\mathbf{X}_i,\mathbf{Y}_i,\mathbf{Z}_i)\mathbf{s}^\top(\hat{\xi}_f,\mathbf{X}_i,\mathbf{Y}_i,\mathbf{Z}_i)|\mathbf{X}_i,\mathbf{Y}_i,\mathbf{Z}_i]$$

where v is the number of samples used to approximate the covariance matrix and $\mathbf{X}_i$, $\mathbf{Y}_i$, and $\mathbf{Z}_i$ is an imputation from $f_\xi(\mathbf{Z}|\mathbf{X},\mathbf{Y})$. The third term in Equation 3.26 is 0 when ξ_f is evaluated at the maximum likelihood estimate $\hat{\xi}_f$. The number of samples used to approximate the covariance matrix needs to be selected. This number tends to be large (e.g. 1000 or more) (Houts & Cai, 2015). To choose the number of samples, I first ran the algorithm

using 1000 samples. I compared the standard error estimates to the Monte Carlo standard deviations of the model parameters and saw severe underestimation. I then increase the number to 2000 and saw improvement in the standard error estimates. I further increased them to 3000 and saw little improvement. Because of the additional computational burden of using 3000 samples and the small benefit in standard error estimation, I set the value at 2000 for all simulation conditions.

Complete Data Models and Derivatives for Steps 2 and 3

In the second step of the MH-RM algorithm, stochastic approximation, Equation 2.26 is approximated as the sample average of complete data gradients and the conditional expected distribution of the missing data, given the observed data are calculated. In the third step, the RM update is made. The complete data log-likelihood $l(\xi_f|\mathbf{Y},\mathbf{X})$ and its derivatives are needed for these two steps. As such, the first and second derivatives of the complete data models with respect to unrestricted parameters are described in the next subsections.

Latent Structure Models

$$l = -\frac{1}{2}[\eta - \bar{}(\xi_f)]'[\Sigma(\tau)]^{-1}[\eta - \bar{}(\xi_f)] - \frac{1}{2}\log|\Sigma(\tau)| - \frac{1}{2}N\log 2\pi. \tag{3.29}$$

The first derivative of l with respect to the parameter vector ξ_f is

$$\frac{\partial l}{\partial \xi_f} = \frac{\partial \bar{}'}{\partial \xi_f}\Sigma(\tau)^{-1}(\eta - \bar{}\xi_f). \tag{3.30}$$

The first derivative of l with respect to a parameter τ_r is

$$\frac{\partial l}{\partial \tau_r} = -\frac{1}{2}\left[tr\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_r}\right) - (\eta - \bar{})'\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_r}\Sigma^{-1}(\eta - \mu)\right]. \tag{3.31}$$

The second derivative of l with respect to ξ_f is

$$\frac{\partial^2 l}{\partial \xi_f \partial \xi_f^{'}} = \frac{\partial \mu^{'}}{\partial \xi_f} \Sigma^{-1} \frac{\partial \mu^{-1}}{\partial \Sigma^{-1}} + (\eta - \mu)^{'} \Sigma^{-1} \frac{\partial^2 \mu}{\partial \xi_f \partial \xi_f^{'}}. \tag{3.32}$$

The second derivative of l with respect to parameter τ_r is

$$\begin{aligned} \frac{\partial^2 l}{\partial \tau_r \partial \tau_r^{'}} = -\frac{1}{2}[tr(\Sigma^{-1} \frac{\partial \Sigma}{\partial \tau_r} \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \tau_r \partial \tau_r^{'}}) \\ + (\eta - \mu)^{'}[(-1)\Sigma^{-1} \frac{\partial \Sigma}{\partial \tau_r} \Sigma^{-1} + \Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \tau_r \partial \tau_r^{'}} \Sigma^{-1} \\ - \Sigma^{-1} \frac{\partial \Sigma}{\partial \tau_r} \Sigma^{-1}](\eta - \mu)]. \end{aligned} \tag{3.33}$$

Measurement Models Recall the 2PL model for dichotomous data (Equation 2.7), suppressing i and k subscripts and the dependency of P on a and c, define

$$P = \frac{\exp(a\theta + c)}{1 + \exp(a\theta + c)} \tag{3.34}$$

The complete data log-likelihood can be written as

$$l = \sum x \log P + (1 - x) \log(1 - P). \tag{3.35}$$

The first derivatives of the complete data log-likelihood are as follows:

$$\begin{aligned} \frac{\partial l}{\partial a} = \frac{\partial l}{\partial P} \frac{\partial P}{\partial a}, \\ \frac{\partial l}{\partial c} = \frac{\partial l}{\partial P} \frac{\partial P}{\partial c} \end{aligned}$$

where

$$\frac{\partial l}{\partial P} = \frac{x-P}{P(1-P)},$$
$$\frac{\partial P}{\partial a} = P(1-P)\theta,$$
$$\frac{\partial P}{\partial c} = P(1-P)$$

The second derivatives are given by the following:

$$\frac{\partial^2 l}{\partial a \partial a} = (\frac{\partial}{\partial a}\frac{\partial l}{\partial P})\frac{\partial P}{\partial a} + \frac{\partial l}{\partial P}(\frac{\partial}{\partial a}\frac{\partial P}{\partial a}),$$
$$\frac{\partial^2 l}{\partial c \partial c} = (\frac{\partial}{\partial c}\frac{\partial l}{\partial P})\frac{\partial P}{\partial c} + \frac{\partial l}{\partial P}(\frac{\partial}{\partial c}\frac{\partial P}{\partial c}),$$
$$\frac{\partial^2 l}{\partial a \partial c} = (\frac{\partial}{\partial a}\frac{\partial l}{\partial P})\frac{\partial P}{\partial c} + \frac{\partial l}{\partial P}(\frac{\partial}{\partial a}\frac{\partial P}{\partial c})$$

where

$$\frac{\partial}{\partial a}\frac{\partial l}{\partial P} = (-\frac{x}{P^2} - \frac{1-x}{(1-P)^2})\frac{\partial P}{\partial a},$$
$$\frac{\partial}{\partial c}\frac{\partial l}{\partial P} = (-\frac{x}{P^2} - \frac{1-x}{(1-P)^2})\frac{\partial P}{\partial c},$$
$$\frac{\partial}{\partial a}\frac{\partial P}{\partial a} = P(1-P)(1-2P)\theta^2,$$
$$\frac{\partial}{\partial c}\frac{\partial P}{\partial c} = P(1-P)(1-2P),$$
$$\frac{\partial}{\partial a}\frac{\partial P}{\partial c} = P(1-P)(1-2P)\theta$$

Chapter 4: Simulation Studies

To examine the performance of the proposed HRD and MRD models, I carried out two Monte Carlo simulation studies. The first study had two purposes: 1) To examine the recovery of measurement and structural parameters under FIML estimation with the MH-RM algorithm, and 2) To explore the proposed models' parameter recovery under two misspecifications. The purpose of the second study is to examine the recovery of several treatment effect estimates under both models with varying conditions: 1) the ATE for participants within 1 point of the cutoff value, $ATE_{c\pm1}$, 2) the range of the LATE for the middle 90% of participants at the cutoff, $[LATE_{q.05}, LATE_{q.95}]$, and 3) the ATE within the bottom 30% of participants in the population of the latent construct, ATE_p. The performance of the proposed models in terms of estimating a LATE with respect to the ORV, $LATE_o$ and with respect to the LRV, $LATE_l$ was compared to the conventional observed variable two-stage estimation approach. This chapter describes the methods and summarizes the results of the two studies.

4.1 Simulation Study I

4.1.1 Purpose

The first Monte Carlo simulation evaluated if implemented R code properly recovered item parameters and structural parameters under ideal conditions. Furthermore, the recovery of model parameters was assessed under two mispecifications of the proposed models

- when the interaction term is omitted and when the covariate is omitted. These misspecifications are a violation of the RD model assumptions, as the error term is now dependent on omitted variables.

4.1.2 Methods

The ideal measurement and sampling conditions were as follows: A sample size of 500 at level-2 was chosen based on pilot simulations with sample sizes between 100 and 1000 as well as the availability of large scale assessment studies such as the Early Childhood Longitudinal Study with over 1000 clusters); a test length of 30 items was used as it has shown to recover 2PL item parameters well under a variety of conditions (Şahin & Anıl, 2017); balanced clusters with 20 individuals per cluster was selected based on typical sample sizes in education research assessing classroom or school traits (Lüdtke et al., 2008); the ICC was set at 0.2 to represent a reasonable magnitude as ICC values are generally smaller than 0.3 in educational research (Bliese, 2000; James, 1982); the data have no missing responses, and appropriate magnitudes of item slopes and intercepts for normal and item responses models (e.g., 2PL model) are used (see Section 4.1.2).

Structural parameters reflect effect sizes seen in the education literature. A minimal detectable effect size (MDES) of 0.25 is typical for large scale educational evaluations (Bloom, Richburg-Hayes, & Black, 2007; Schochet, 2008), and guidance on RD power calculations commonly use MDES values between 0.2 and 0.5 (see e.g., H. Lee & Munk, 2008; Rhoads & Dye, 2016). As such, the value of the LATE will be set at 0.45 . Based on RD studies in education (Jung, 2010; Luyten, 2006), the following values will be used for the HRD model described in Section 3.1.2: $\gamma_{00} = 0$, $\beta_1 = 0.50$, $\beta_2 = 0.50$, $\beta_3 = 0.10$, $\beta_4 = 0.15$, $\tau_0^2 = 0.25$. For the MRD model (Section 3.1.2) the following values will be used: $\gamma_{00} = 0$, $\beta_1 = 0.50$, $\gamma_{20} = 0.50$, $\beta_3 = 0.10$, $\beta_4 = 0.15$, $\tau_0^2 = 0.25$, $\tau_{20} = 0.25$, and $\tau_2^2 = 0$. Note, the

$\tau_{20} = 0$ indicates there is no relationship between the intercept (i.e., the average outcome for the control group after controlling for the effect of the RV) and the treatment effect. RD analysis will be conducted using the full dataset a bandwidth of 1.0.

Data Generation

Data generation comprises three steps.

Step 1: Generating item parameters and latent variables The slope parameters for the latent predictors and latent outcome were drawn from a lognormal distribution with a mean of 0.3 and a standard deviation of 0.20 (on the normal scale), truncated to the interval [1.0, 2.5]. Similarly, the intercept parameters were generated from a standard normal distribution truncated to [-2, 2] (Feinberg & Rubright, 2016; Mislevy & Stocking, 1989). The item parameters are treated as fixed across all conditions, per test length. Within each replication, the latent running variable, $\theta_{r,ij}$ the latent covariate $\theta_{c,ij}$ are computed as the sum of the latent group means, drawn

$$\theta_j \sim N_2(\mathbf{0}, \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}),$$

and the individual deviations from the group means, drawn

$$\delta_{ij} \sim N_2(\mathbf{0}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}).$$

The latent outcome, η_{ij} is calculated using Equation 3.8 and 3.11 once the random variables have been generated and fixed effects have been specified.

Step 2: Generating observed indicators Item responses are generated independently for each latent variable. Item parameters and latent variable values are used in Equation 3.2

to calculate probability of endorsement of each item for each individual. These probabilities are compared to draws from a uniform distribution $[0,1]$ to generate item responses.

Step 3. Participant classification Participants are classified based on the ORV. In practice the cutoff either naturally exists or is chosen based on the research question. In the current study, the cutoff was chosen so that approximately 30% of the data falls below the cutoff. In order to select cutoff values that reflect the 30th percentile in the population, I generated large samples to approximate the population. First, I generated $10,000,000$ sets of item responses for the RV, as described above, and converted them into summed scores. For the MRD model I then set the $30th$ percentile of the distribution of the summed scores as the cutoff value, $c = 10$. For the HRD model, I then aggregated the summed scores into $100,000$ clusters by calculating the mean; the $30th$ percentile was set as the cutoff, $c = 13.17$.

Under the proposed model, participants are classified based on their ORV scores. Accordingly, the generated item responses for the LRV were summed to create ORV scores for each participant. The ORV scores were compared to the cutoff value. Participants were assigned to treatment if their scores were at or below the cutoff and the control group if their scores were above the cutoff.

Model For study I, the true data generating model was fitted to the simulated data using FIML via the MH-RM algorithm. Model equations can be found in Sections 3.1.2 and 3.1.2. Standard errors were calculated via direct application of the Louis formula after the MH-RM algorithm converged (see Section 2.5.3). To assess impact of model misspecification, a violation of one of the latent RD assumptions (see Section 3.2.1), the HRD and MRD models will be fit to the full data under two misspecified conditions: 1) omitting the interaction term, $\theta_{r,ij}T_{ij}$ and its corresponding regression parameter, β_3 and 2) omitting the latent covariate, $\theta_{c,ij}$ and its corresponding regression parameter, β_4.

Bandwidths The models were fit to the full data as well as partial data defined by a

bandwidth of 1.0. The use of bandwidths is more complex with a LV RD model than with a conventional RD model, as the measurement model would be misspecified if the full data is not used, a violation of one of the LV RD model's assumptions. In order to avoid this issue, I first fit the full data to the model in order to estimate item parameters under a correctly specified measurement model. These item parameters, estimated from the full data, were then used when fitting the proposed models using a bandwidth; consequently, item parameters were not estimated from the partial data and the model can still be fit when using a bandwidth. The ORV scores (i.e., summed scores) were used to identify the data that falls within the bandwidth. An interval was calculated around the cutoff value such that, $[c \pm SD(ORV) \times 1.0]$, where $SD(ORV)$ is the standard deviation of the ORV scores and 1.0 is the bandwidth value. Only the individuals with ORV scores within the interval were included in the analysis. The proposed models were then fit to the partial data.

Evaluation Five hundred replications were completed. The measurement and structural parameter estimates and LATE were evaluated in terms of their absolute and relative bias as well as root mean squared error (RMSE). Estimated standard errors for the model parameters and the LATE were compared against the Monte Carlo standard deviations. 95% confidence intervals for the structural parameters are also evaluated by the empirical coverage.

4.1.3 MHRM Algorithm: Additional Considerations and Convergence

Before the MHRM algorithm can be implemented, the sampler must be tuned and convergence criteria must be set.

Tuning Constants

As described in Section3.2.1, the distribution of the level-1 and level-2 missing data are defined by μ_{1j}, Σ_{1j}, μ_2, and Σ_2. The random variables, which are treated as missing data, at level 1 are δ_{ij}, the individual deviations from cluster means for the latent predictors and ε_{ij}, the level-1 error terms. As the clustered means are centered on the grand mean and the level-1 residuals follow a standard normal distribution with a variance of 1, $\mu_{1j} = (0,0,0)'$ and Σ_{1j} is an identity matrix in both models. Similarly, θ_j, the cluster means of the latent predictors and $\mathbf{u}_j$ are treated as missing data at level 2. The means of these random components are assumed to be zero, making $\mu_2 = (0,0,0)'$ in the HRD model and $\mu_2 = (0,0,0,0)'$ in the MRD model. The variance-covariance matrix at level 2 is as described in Section 3.1.2 and Section 3.1.2. The values of μ_{1j}, Σ_{1j}, μ_2, and Σ_2 are summarized in Table 4.1.

As discussed in Section 3.2.1, the desired acceptance rate in the MH sampler is achieved by adjusting the value of the tuning constant *w*. In order to determine the values of *w* which yield acceptance rates between the target rates of 20 and 30% at each level (Gelman, Gilks, & Roberts, 1997), I explored the acceptance rates using various tuning constants using a sample size of 4000 with 200 clusters. The results were similar across models, but varied by test length (see Table 4.2 and Table 4.3).

The values of *w* were further tuned for each simulation condition, resulting in values between 0.60 and 0.25 at level-1 and 0.05 and 0.02 at level-2 for the HRD model and values between 0.55 and 0.28 at level-1 and 0.06 and 0.01 at level-2 for the MRD model. Notably, the MRD model was more sensitive to values of *w* across conditions, requiring changes in the hundredth decimal place across to achieve appropriate acceptance rates, while one set of *w* values was successfully used for all 10-item and one set for all 30-item conditions under the HRD model.

Table 4.1: Latent Variable Distributions for Multilevel Latent Regression Discontinuity Models

Hierarchical Regression Discontinuity Model				
Level 1	$\begin{bmatrix}\delta_{r,ij}\\ \delta_{c,ij}\\ \varepsilon_{ij}\end{bmatrix}$	$\sim \quad \mu_{1j} =$	$\begin{bmatrix}0\\0\\0\end{bmatrix}$	$\Sigma_{1j} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$
Level 2	$\begin{bmatrix}\theta_{r,j}\\ \theta_{c,j}\\ u_{0j}\end{bmatrix}$	$\sim \quad \mu_{2} =$	$\begin{bmatrix}0\\0\\0\end{bmatrix}$	$\Sigma_{2} = \begin{bmatrix}\text{Var}(\theta_{r,j}) & 0 & 0\\ 0 & \text{Var}(\theta_{c,j}) & 0\\ 0 & 0 & \tau_0^2\end{bmatrix}$
Multisite Discontinuity Model				
Level 1	$\begin{bmatrix}\delta_{r,ij}\\ \delta_{c,ij}\\ \varepsilon_{ij}\end{bmatrix}$	$\sim \quad \mu_{1j} =$	$\begin{bmatrix}0\\0\\0\end{bmatrix}$	$\Sigma_{1j} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}$
Level 2	$\begin{bmatrix}\theta_{r,j}\\ \theta_{c,j}\\ u_{0j}\\ u_{2j}\end{bmatrix}$	$\sim \quad \mu_{2} =$	$\begin{bmatrix}0\\0\\0\\0\end{bmatrix}$	$\Sigma_{2} = \begin{bmatrix}\text{Var}(\theta_{r,j}) & 0 & 0 & 0\\ 0 & \text{Var}(\theta_{c,j}) & 0 & 0\\ 0 & 0 & \tau_0^2 & \tau_{20}\\ 0 & 0 & \tau_{20} & \tau_2^2\end{bmatrix}$

Table 4.2: Tuning Constants for Multilevel Latent RD Models with 10 Items

		Set 1		Set 2		Set 3		Set 4	
		L1	L2	L1	L2	L1	L2	L1	L2
HRD	w	1.0	1.0	0.75	0.3	0.5	0.1	0.6	0.05
	AR (%)	17 - 20	1 - 3	22 - 26	3 - 6	29 - 33	10 - 13	27-30	22 - 25
RD	w	1.0	1.0	0.75	0.3	0.5	0.1	0.6	0.05
	AR (%)	16 - 19	0.4 - 0.7	23 - 26	2 - 4	30 - 34	7 - 9	25 - 29	20 - 23

Table 4.3: Tuning Constants for Multilevel Latent RD Models with 30 Items

		Set 1		Set 2		Set 3		Set 4	
		L1	L2	L1	L2	L1	L2	L1	L2
HRD	w	1.0	1.0	0.75	0.3	0.5	0.1	0.25	0.01
	AR (%)	5 - 8	0 - 1	8 - 11	0 - 2	14 - 16	1 - 4	25 - 29	29 - 32
MRD	w	1.0	1.0	0.75	0.3	0.5	0.1	0.25	0.01
	AR (%)	6 - 9	0 - 2	9 - 11	1 - 3	13 - 16	3 - 7	25 - 28	30 - 33

"Burn-In"

The number of "burn-in" cycles for the MH sampler must also be determined; I did this in two steps. First, I examined the auto correlations of random drawings under a simulated condition for each model: 4000 individuals nested in 200 clusters using a 20-item test length. Time series plots for 5 randomly selected from each of the random effects at both levels appear in the Appendix. These plots suggest at least 20 burn-in cycles is reasonable for the HRD model about 50 burn-in cycles for the MRD model. In the next step, I ran the full MH-RM algorithm, and examined the trace plots of the parameter estimates. Noting the long run time, I also examined the trace plots using 5 burn-in cycles for the HRD model and 10 burn-in cycles for the MRD model, which appear in the Appendix. Because the trace plots of the parameter estimates look similar under both conditions, 5 and 10 burn-in cycles were used for the HRD and MRD models, respectively, in the study.

Starting Values for Stage 1

The MH-RM algorithm requires starting values for all model parameters. Standardized summed scores were used as starting values for the LVs, $\theta_{r,ij}$, $\theta_{c,ij}$, and η_{ij}. These scaled summed scores were used in a conventional multilevel RD model; the estimates from this model were used as the starting values for the the fixed effects. Starting values for the level-1 residuals were generated from a standard normal distribution. Similarly, starting values for the level-2 residuals, u_{0j} and u_{2j} as well as the cluster-level latent predictors, $\theta_{r,j}$ and $\theta_{c,j}$ were each generated from a $N(0, 0.25^2)$ distribution.

Convergence

The MH-RM algorithm employs the use of adaptive gain constants in a three-stage procedure in order to avoid premature convergence (Cai, 2008). In the first stage, $M1$, iterations are run with the gain constant fixed at 1. In the second stage, $M2$, further iterations are run; the parameter estimates calculated during this second stage are averaged and used as starting values in the third stage, $M3$, during which a decreasing gain constant is employed. In order to check convergence, Cai (2008) proposed to monitor a "window" of the difference between successive parameter estimates. The algorithm convergences once the largest difference in the window, suggest to be 3 consecutive differences (Cai, 2008), is below a certain number.

Appropriate gain constants, number of iterations, and convergence criteria is assessed by examining trace plots with a large number of iterations (e.g., 1000 for $M1$ and $M2$ combined and 1000 for $M3$) for parameter estimates. The Appendix shows trace plots of measurement and structural parameters when the constant gain was 0.2 and the decreasing gain constant, g_t at the tth iteration is defined as

$$g_t = \frac{0.2}{t^{\varepsilon}} \tag{4.1}$$

The value of ε was set at 0.75 after examining the trace plots of model parameter estimates.

The trace plots suggest parameter estimates tend to move close to the MLEs within 100 iterations under both models; estimates are oscillating around the MLEs within 200 to 400 iterations for the HRD model and within 300 to 500 iterations for the MRD model. When 1.0×10^{-5} was used as convergence criteria both models converged within 300 $M3$ iterations and parameter estimates were close to their true values. As such, $M1$ was set at 100 for both models, $M2$ was set at 300 for the HRD model and 500 for the MRD model,

Table 4.4: Gradient Norms for Simulation Study I Conditions

Sample	HRD Model		MRD Model	
	Average	Range	Average	Range
Full Sample	<0.01	0 - 0.02	<0.01	0 - 0.02
Bandwidth	<0.01	<0.01 - 0.02	<0.01	0 - 0.03

and $M3$ was set at 400 for both models.

Furthermore, in order to assess whether the algorithm has converged at a local maximum, I check whether the gradient is sufficiently close to 0 and that the Hessian information is positive definite. Finally, I compute the condition number of the information matrix.

4.1.4 Results

HRD and MRD Models using the Full Sample

The HRD and MRD models were fit to the generated data. The marginal reliability (Thissen & Orlando, 2001), which is the reliability of the LRV (i.e., the IRT scores calculated for the LRV), under both generating models was approximately 0.90 and the Cronbach's alpha is 0.88. All replications converged (i.e., positive definite observed data information matrix) under both models. The condition numbers of the information matrix ranged from 109 to 185 under the HRD model and from 176 to 219 under the MRD model. The average runtime for the HRD and MRD models, was 14 and 23 minutes, respectively. To assess whether the algorithm converged at a local maximum, I checked whether the gradient was sufficiently close to 0. The results appear in Table 4.4.

The measurement and structural parameters were well recovered under the HRD and MRD models, though estimates of the regression coefficients showed slightly more bias under the MRD model. The recovery of these measurement parameters was similar across both models (see Figure 4.1). The true generating item slope and item intercept values and

corresponding bias, relative bias, and RMSE for the RV, covariate, and outcome variable for both models appear in the Appendix. The relative bias in the item parameter estimates is generally below 2% for the slopes and below 5% for the intercepts with the exception of the outcome variable which has several intercept estimates with relative bias values above 20%. The item slope estimates tend to be more stable with RMSE values below .05, than the item intercept estimates, with RMSE values above .05.

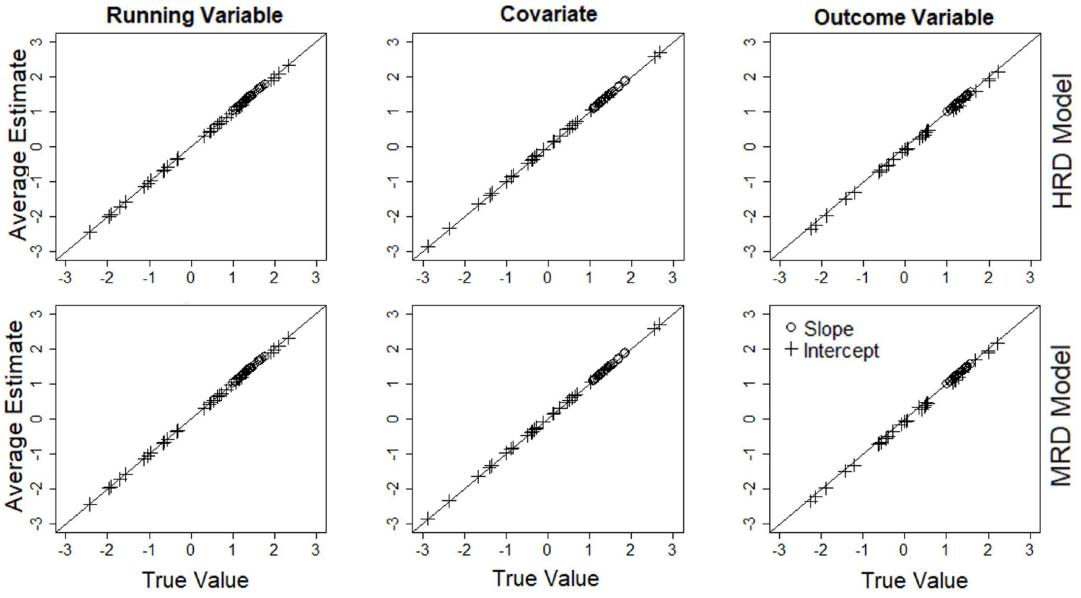

Figure 4.1: True Generating Item Parameters Plotted against Average Estimate over 500 Replications using the Full Sample

The true generating structural parameters and their corresponding bias, relative bias, and RMSE for both models appear in Table 4.5. Estimates for the regression coefficients tended to be better under the HRD model than under the MRD model. The estimate for β_3, the interaction between the treatment assignment and the RV, tended to be the most biased as it was over estimated by between 16 and 19% across models, while the estimate for β_2 and γ_{20}, the effect of treatment assignment, tended to be the least precise with an RMSE of .07 and .10 in the HRD and MRD models, respectively. The variance components tended

Table 4.5: Bias, Relative Bias, and RMSE for Structural Parameters and LATE at the Latent Cutoff for Simulation Study I with Full Sample

Par	True Value	HRD			MRD		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
β_1	0.50	-0.01	-1%	0.03	-0.02	-3%	0.04
β_2	0.50	<0.01	<1%	0.07	NA	NA	NA
γ_{20}	0.50	NA	NA	NA	-0.02	-4%	0.10
β_3	0.10	0.02	16%	0.05	0.02	19%	0.09
β_4	0.25	<0.01	<1%	0.02	<0.01	1%	0.03
$Var(\theta_{r,j})$	0.25	-0.02	-10%	0.03	<0.01	<1%	0.04
$Var(\theta_{c,j})$	0.25	-0.02	-7%	0.03	<0.01	1%	0.03
τ_0	0.25	<0.01	-2%	0.02	<0.01	2%	0.04
τ_2	0.55	NA	NA	NA	-0.01	-2%	0.06
τ_{20}	0.00	NA	NA	NA	<0.01	NA	0.03
$LATE_l$	0.45	-.02	-4%	0.05	-0.04	-8%	0.09

to be underestimated in the HRD model with relative bias values between -2 and -10% and well recovered under the MRD model with relative bias values no more extreme than 2%. As with the regression coefficients, the LATE with respect to the LRV, $LATE_l$ is slightly underestimated under both models with a relative bias of -4% in the HRD model and -8% in the MRD model. The $LATE_l$ also had a smaller RMSE under the HRD model, .05, than under the MRD model .09.

The post-convergence estimation of standard errors resulted in generally underestimated standard errors in both models. Figure 4.2 shows the average standard errors plotted against the Monte Carlo standard deviation for slope and intercept parameters for the three latent variables. The largest difference in the standard error estimates and Monte Carlo standard deviations is 0.04 in the item slopes and 0.02 in the item intercepts. Standard errors were also generally underestimated for the structural parameters (see Figures 4.3 and 4.4), with the most severe underestimation for, τ_2^2 the variance of the level-2 random effects, u_{2j}.

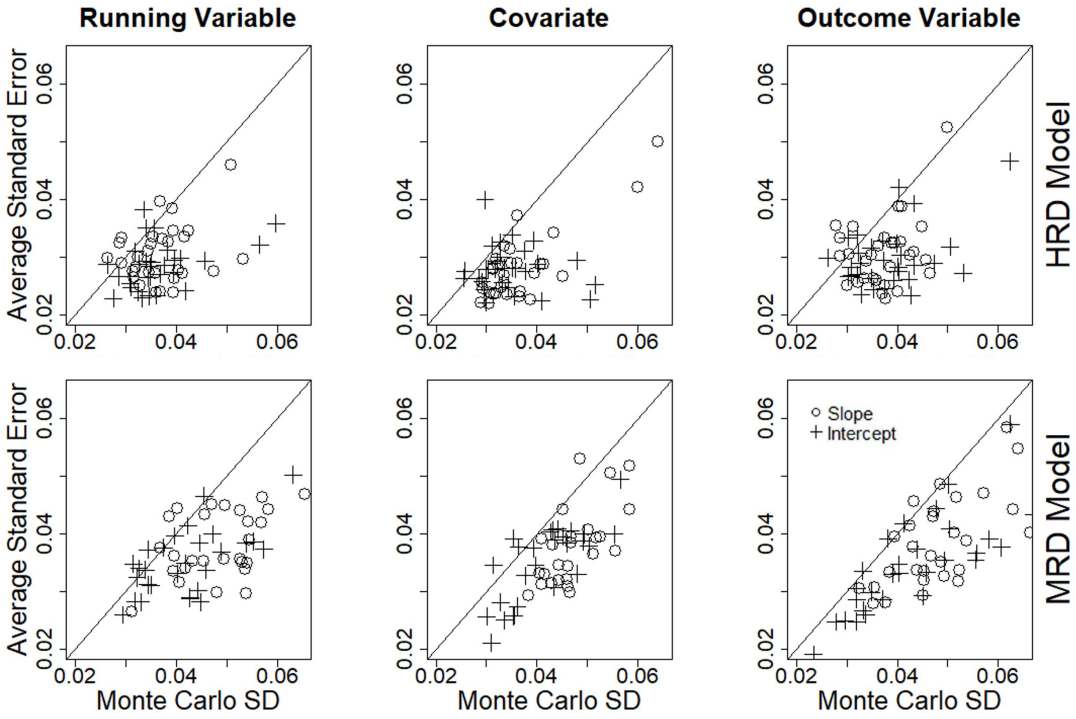

Figure 4.2: Average Standard Error Estimates Plotted against Monte Carlo Standard Deviations for Measurement Parameters for Simulation Study I with Full Data

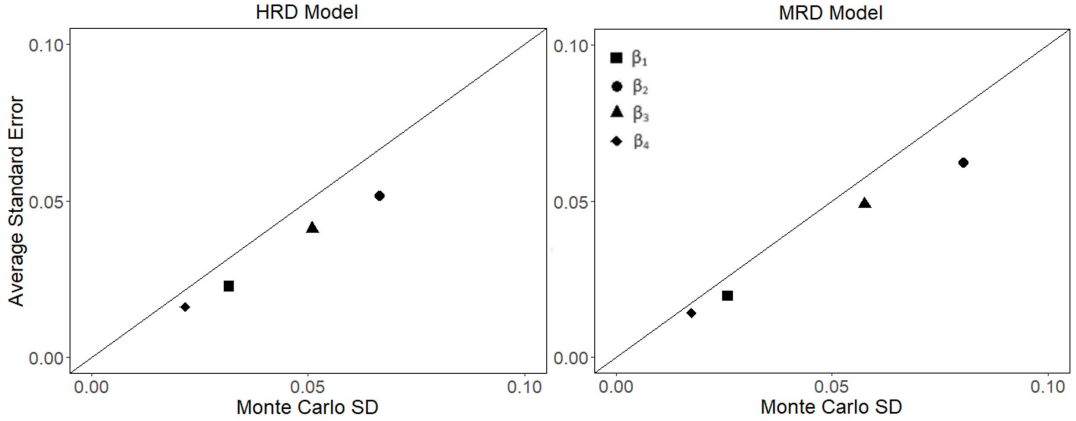

Figure 4.3: Average Standard Error Estimates Plotted against Monte Carlo Standard Deviations for Regression Parameters for Simulation Study I with Full Data

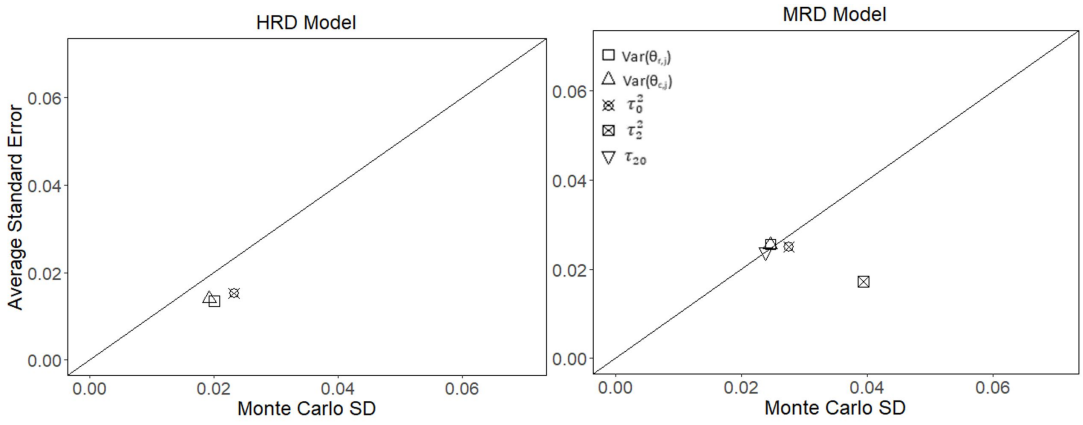

Figure 4.4: Average Standard Error Estimates Plotted against Monte Carlo Standard Deviations for Variance Parameters for Simulation Study I with Full Data

The 95% interval coverage for the regression parameters is near the nominal level when using the full sample under both models, with coverage above 89% (see Figure 4.5. The coverage of the variance parameters tends to be slightly lower under the HRD model than the MRD model, especially for the variance of the cluster level RV, $Var(\theta_{r,j})$ (see Figure 4.6).

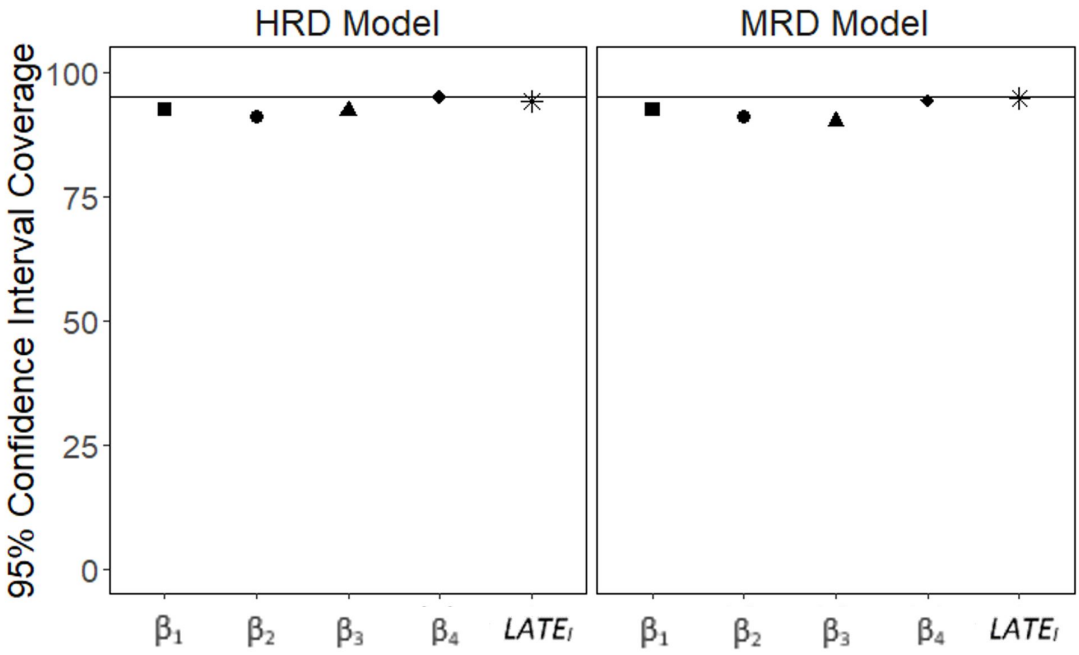

Figure 4.5: 95% Confidence Interval Coverage of Regression Parameters and LATE for HRD and MRD Models using the full sample

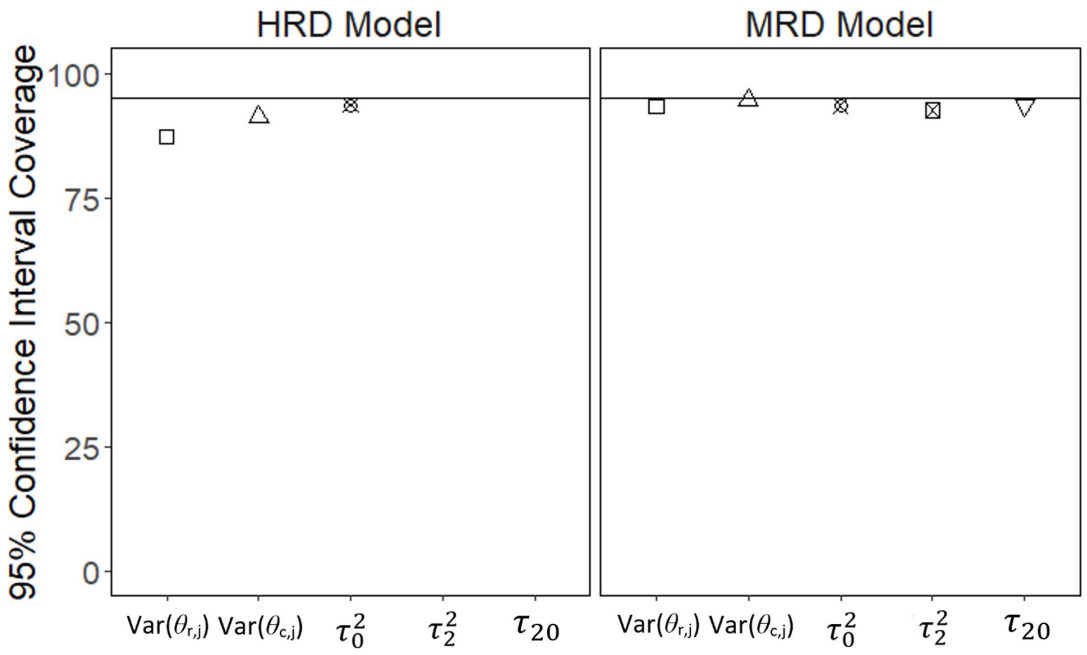

Figure 4.6: 95% Confidence Interval Coverage of Variance Parameters for HRD and MRD Models using the full sample

HRD and MRD Models using a Bandwidth

As described in Section 4.1, when using a bandwidth, the full data were first used to estimate the item parameters. The data was then subset so that only participants within 1 standard deviation of the cutoff (i.e., bandwidth = 1) were included in the sample. The model was then estimated again using only the subset of the data with the item parameters that had been previously estimated fixed. All replications properly converged under both models. The range of condition numbers of the information matrix for the replications was 439 to 868 for the HRD model and 538 to 1399 for the MRD model. The average runtime for the HRD and MRD models using a bandwidth was 13 and 18 minutes, respectively. As expected, the measurement parameters were properly recovered under both models with bias and RMSE values similar to those from the full sample conditions. Figure 4.7 shows

the average estimates against the true generating values for all the measurement parameters under the HRD and MRD models when using the a bandwidth. The standard error recovery for the measurement parameters when using a bandwidth is also comparable to that when using the full sample.

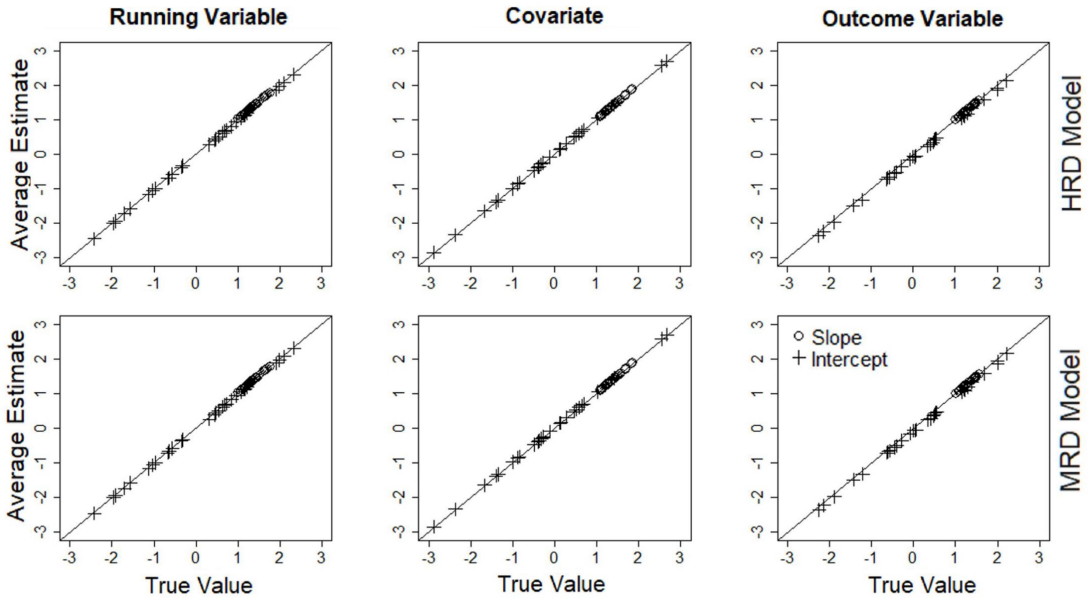

Figure 4.7: True Generating Item Parameters Plotted against Average Estimate over 500 Replications using a Bandwidth

The true generating values and the corresponding bias, relative bias, and RMSE values for the structural parameters appear in the Appendix. The regression parameter estimates are well recovered under the HRD model with relative bias and RMSE values comparable to those from the full model. However, the regression parameter estimates for the relation between the RV and the outcome, β_1 and for the interaction between treatment assignment and the RV, β_3 have large bias and RMSE values. The β_1 parameter is under estimated, relative bias is -48%, while β_3 is over estimated with a relative bias of 140%. The variance parameters for the group level RV, $Var(\theta_{r,j})$, is also severely underestimated in both models

Table 4.6: Bias, Relative Bias, and RMSE for Structural Parameters and LATE at the Latent Cutoff for Simulation Study I with Bandwidth

Par	True Value	HRD			MRD		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
β_1	0.50	0.01	2%	0.04	-0.25	-48%	0.28
β_2	0.50	-0.02	-3%	0.08	NA	NA	NA
γ_{20}	0.50	NA	NA	NA	0.06	12%	0.11
β_3	0.10	0.01	10%	0.06	0.14	140%	0.30
β_4	0.25	<0.01	1%	0.03	<0.01	1%	0.03
$Var(\theta_{r,j})$	0.25	-0.21	-85%	0.21	-0.21	-83%	0.21
$Var(\theta_{c,j})$	0.25	-0.03	-12%	0.04	-0.04	-14%	0.04
τ_0	0.25	-0.01	-6%	0.03	<0.01	-2%	0.03
τ_2	0.55	NA	NA	NA	0.01	-3%	0.04
τ_{20}	0.00	NA	NA	NA	<0.01	NA	0.02
$LATE_l$	0.45	-0.02	-8%	0.08	0.02	4%	0.12

with relative bias values around 84% and RMSE of .21. However, the LATE with respect to the LRV, $LATE_l$ is still well recovered with a relative bias of -8% and RMSE of .08 in the HRD model and relative bias of 4% and RMSE of .12 in the MRD model.

The post-convergence estimation of standard errors resulted in generally underestimated standard errors in both models. As stated above, standard error estimates for the measurement parameters were similar with a bandwidth and when using the full sample. When using a bandwidth, the largest difference in the standard error estimates and Monte Carlo standard deviations is 0.04 in the item slopes and 0.03 in the item intercepts. Standard errors were also generally underestimated for the structural parameters (see Figures 4.8 and 4.9).

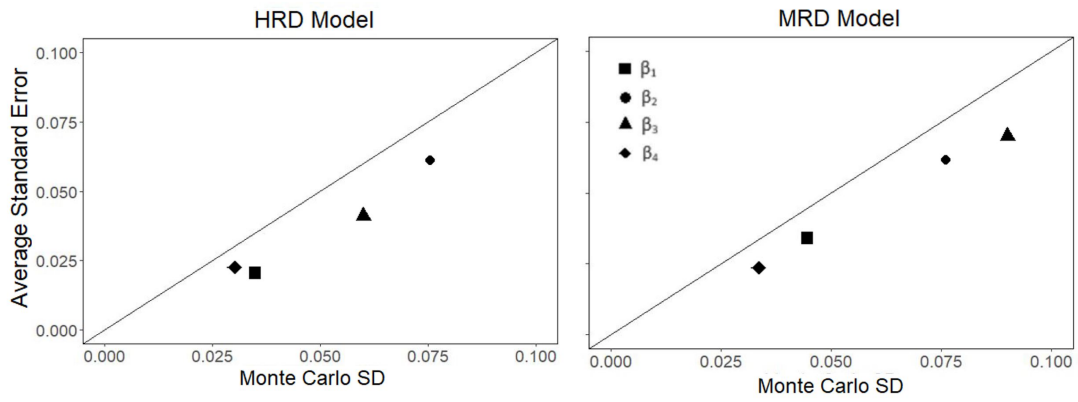

Figure 4.8: Average standard error values of regression coefficients plotted against Monte Carlo standard deviation for HRD and MRD models using a bandwidth of 1.0.

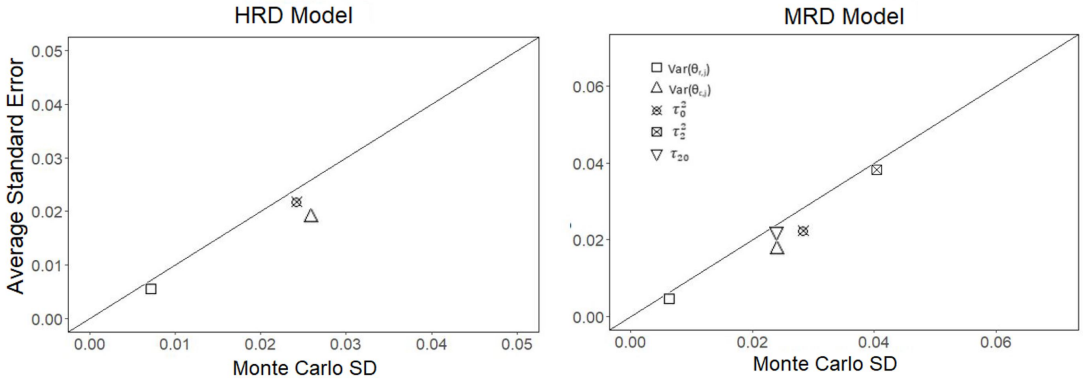

Figure 4.9: Average standard error values of variance parameters plotted against Monte Carlo standard deviation for HRD and MRD models using a bandwidth of 1.0.

The 95% interval coverage for the regression parameters is below the nominal level when using a bandwidth of 1.0 under both models. Coverage is especially poor under the MRD model where the regression coefficient for the relation between the covariate and the outcome is the only regression parameter with coverage above 25%. However, coverage for the $LATE_l$ remains near the nominal level with coverage around 90% under both models (see Figure 4.10). The coverage of the variance parameters is poor for the variance of

the latent predictors at the cluster level (see Figure 4.11), with the coverage being 0% for $Var(\theta_{r,j})$ due to the extreme underestimation of this parameter when using a bandwidth. The coverage of the other variance parameters are greater than 90% in both models.

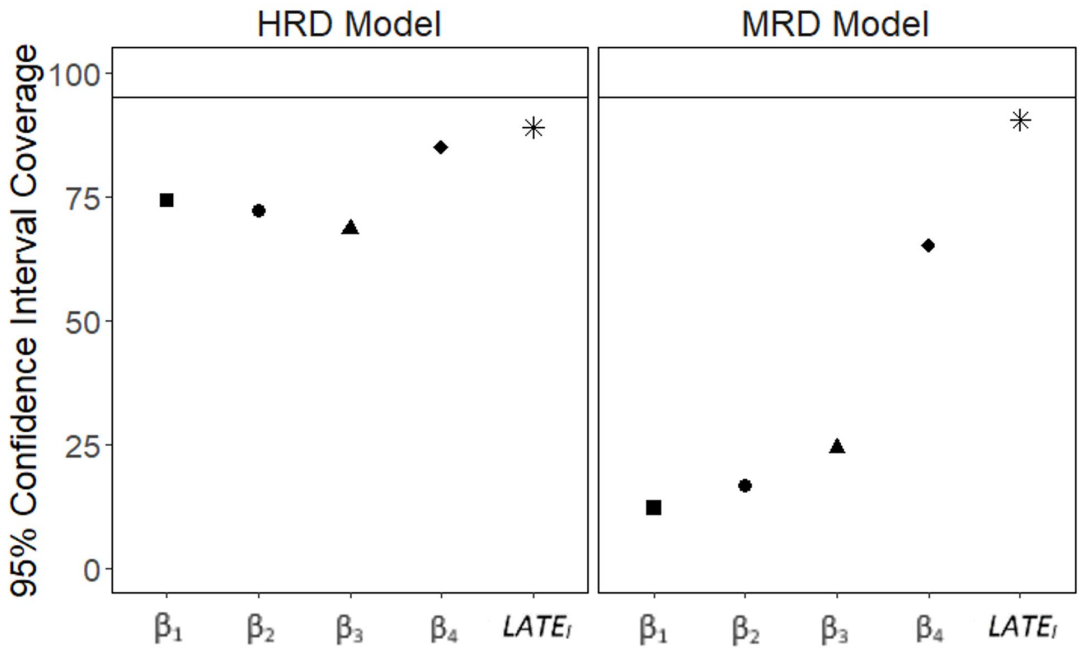

Figure 4.10: 95% Confidence Interval Coverage of Regression Parameters and LATE for HRD and MRD Models using a bandwidth

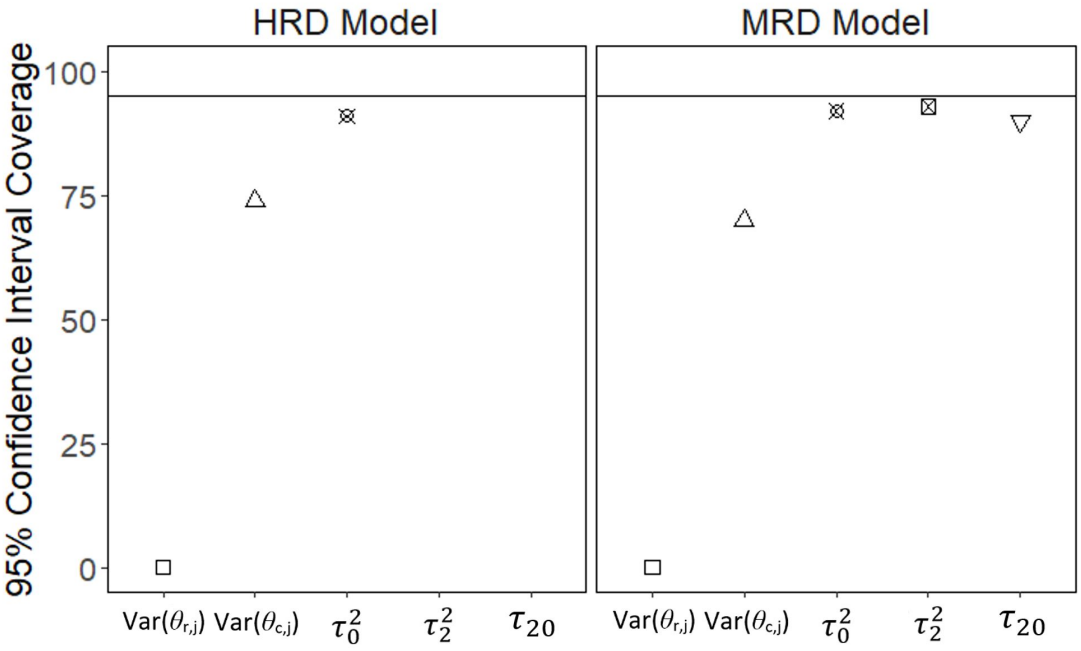

Figure 4.11: 95% Confidence Interval Coverage of Variance Parameters for HRD and MRD Models using a bandwidth

Misspecified Models

No interaction term. One of the assumptions of the latent RD model is that the model is correctly specified. As a preliminary step to assess the impact of violating this assumption, the full sample was fit to each misspecified model, omitting the interaction term. All replications converged for both models. The measurement parameter estimates plotted against the true generating values appear in Figure 4.12. All item parameters are well recovered with the item slopes and intercepts being slightly overestimated for the outcome variable under both models.

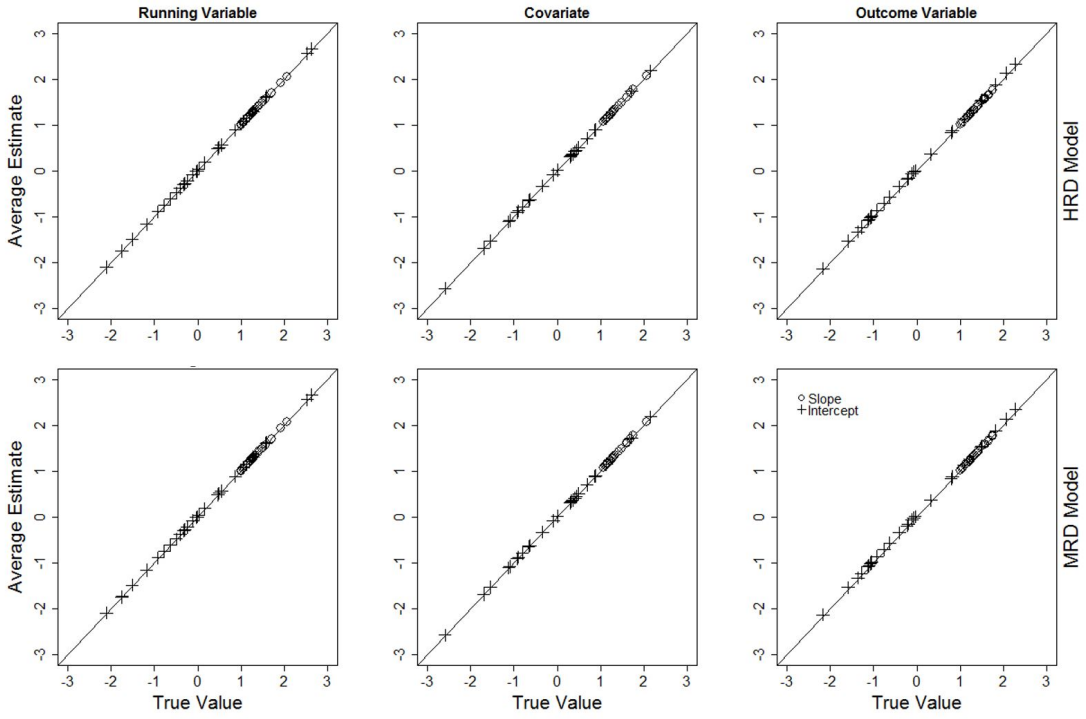

Figure 4.12: Average item parameters plotted against true generating values over 500 replications for misspecified HRD and MRD models with no interaction term using the full sample.

The true generating values for all structural parameters and the $LATE_l$ appear in Table 4.7. The regression parameters are generally well recovered under both misspecified models with the treatment effect parameters, β_2 and γ_{20}, being underestimated more severely under the MRD model, relative bias is -29% and RMSE is .15, than in the HRD model, relative bias is -14% and RMSE is .09. Similarly, the variance estimates are well recovered in the misspecified HRD model but underestimated in the misspecified MRD model, where the relative bias of τ_2 is -22%. As these models were fit without an interaction term, the $LATE_l$ is simply the value of β_2 and γ_{20}. The $LATE_l$ under the misspecified HRD model is slightly underestimated with a relative bias of -5% and moderately underestimated under the misspecified MRD model with a relative bias of -21%.

Table 4.7: Bias, Relative Bias, and RMSE for Structural Parameters and LATE at the Latent Cutoff for Misspecified Models with no Interaction Term using Full Sample

Par	True Value	HRD			MRD		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
β_1	0.50	0.02	5%	0.03	<0.01	<1%	0.03
β_2	0.50	-0.07	-14%	0.09	NA	NA	NA
γ_{20}	0.50	NA	NA	NA	-0.15	-29%	0.15
β_4	0.25	<0.01	1%	0.02	<0.01	<1%	0.02
$Var(\theta_{r,j})$	0.25	-0.01	-2%	0.02	-0.01	-4%	0.03
$Var(\theta_{c,j})$	0.25	<0.01	1%	0.02	-0.01	-3%	0.02
τ_0	0.25	<0.01	<1%	0.02	-0.01	-5%	0.03
τ_2	0.55	NA	NA	NA	-0.06	-22%	0.06
τ_{20}	0.00	NA	NA	NA	0.03	NA	0.04
$LATE_l$	0.45	-.02	-5%	0.09	-0.09	-21%	0.15

The residuals are plotted against the LRV scores in Figure 4.13. There appears to be some heteroskedasticity, especially at the tails with a more narrow spread of residuals at lower LRV values and a broader spread at higher LRV values.

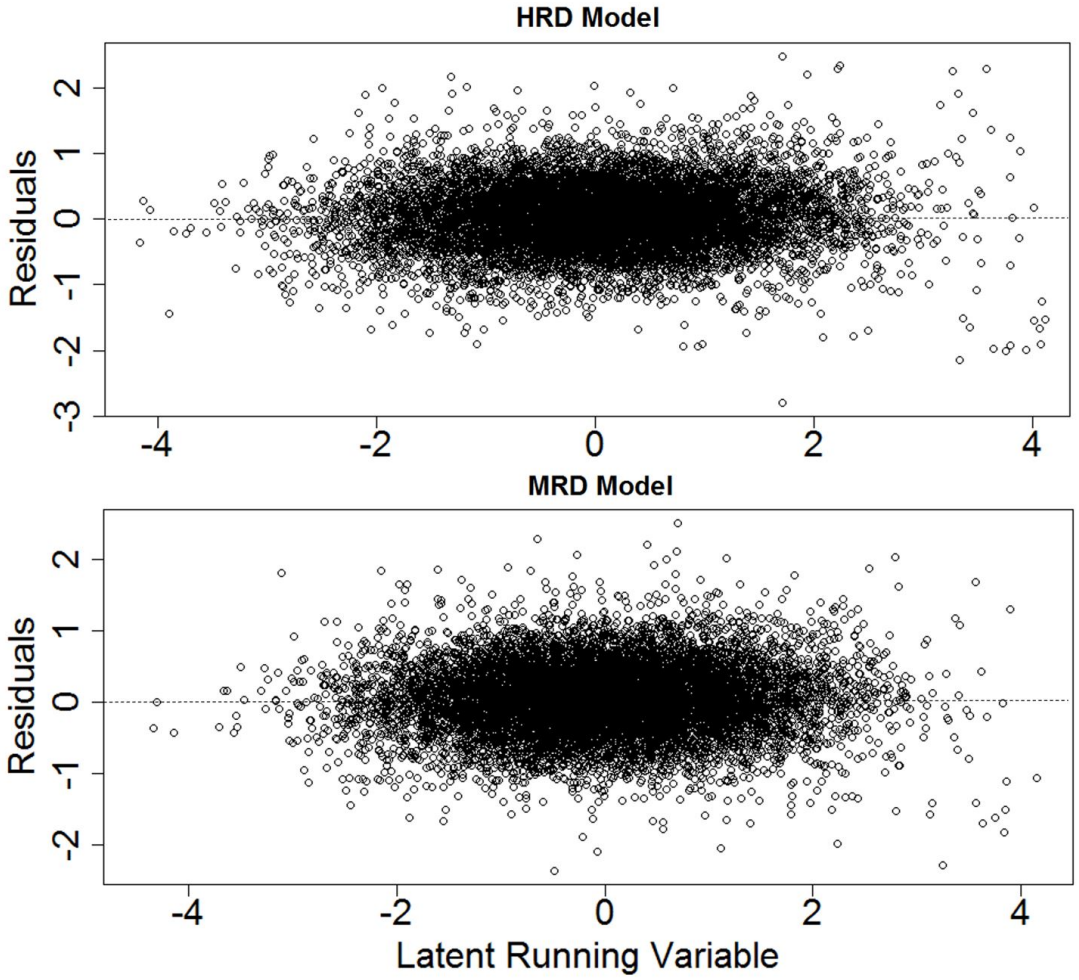

Figure 4.13: Residuals plotted against the latent running variable for the misspecified HRD and MRD models without the interaction term using the full sample.

No Latent Covariate. The full sample was also fit to the misspecified HRD and MRD models, without the latent covariate. All replications converged for both models. The measurement parameter estimates plotted against the true generating values appear in Figure 4.14. The item parameters for the LRV are well recovered under both models. The item slopes tend to be overestimated for the outcome variable under both models, and the outcome variable item intercepts are slightly over estimated under the MRD model.

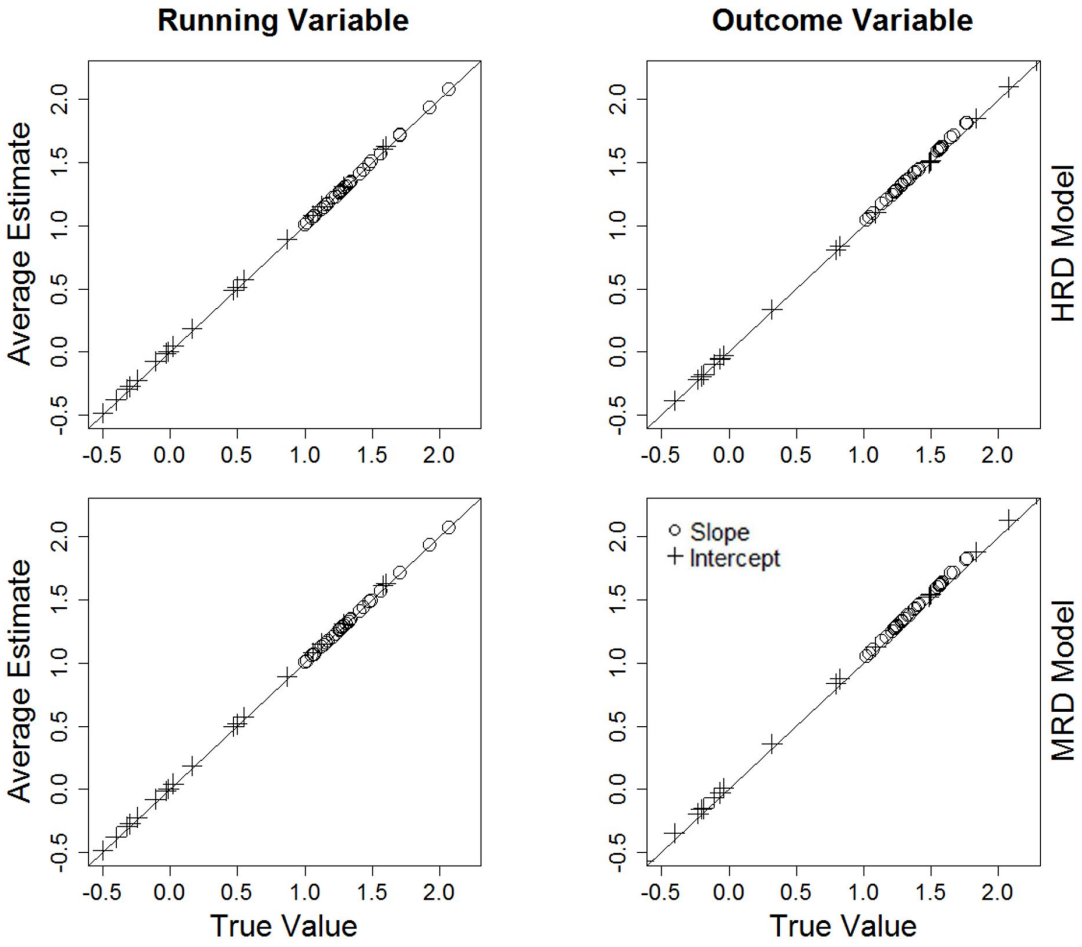

Figure 4.14: Average item parameters plotted against true generating values over 200 replications for misspecified HRD and MRD models with no latent covariate using the full sample.

The true generating values for all structural parameters and the $LATE_l$ appear in Table 4.8. The regression parameters are generally well recovered under both misspecified models with the regression parameter for the interaction term, β_3, being overestimated under the MRD model, relative bias is 20% and RMSE is .08. Similarly, the variance estimates are well recovered in the misspecified HRD model but underestimated in the misspecified MRD model, where the relative bias of τ_2 is -2%. The $LATE_l$ under the misspecified

Table 4.8: Bias, Relative Bias, and RMSE for Structural Parameters and LATE at the Latent Cutoff for Misspecified Models with no Latent Covariate using Full Sample

Par	True Value	HRD			MRD		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
β_1	0.50	-0.01	-3%	0.03	<0.03	-6%	0.05
β_2	0.50	-0.01	-2%	0.07	NA	NA	NA
γ_{20}	0.50	NA	NA	NA	-0.03	-6%	0.10
β_3	0.10	<0.01	<1%	0.05	0.02	20%	0.08
$Var(\theta_{r,j})$	0.25	-0.01	-5%	0.02	<0.01	-1%	0.02
τ_0	0.25	<0.01	1%	0.02	-0.01	-4%	0.02
τ_2	0.55	NA	NA	NA	-0.10	-20%	0.06
τ_{20}	0.00	NA	NA	NA	0.02	NA	0.03
$LATE_l$	0.45	-.01	-2%	0.07	-0.04	-9%	0.08

HRD model is well recovered with a relative bias of -2% and RMSE of .02; it is slightly underestimated under the misspecified MRD model with a relative bias of -9% and RMSE of .08.

The residuals for the two models are plotted against the LRV in Figure 4.15. The residuals tend to be negative at lower values of the LRV and positive at higher values under both misspecified models.

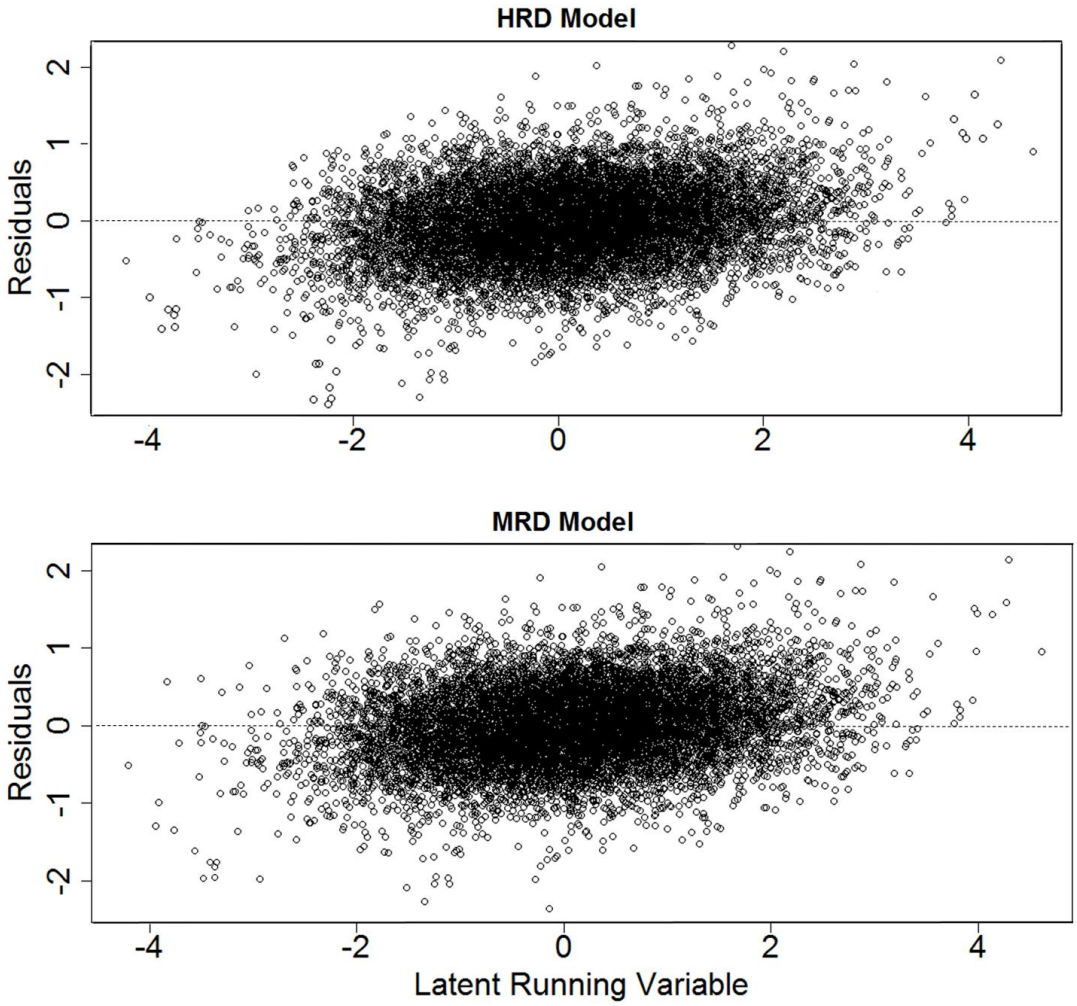

Figure 4.15: Residuals plotted against the latent running variable for the misspecified HRD and MRD models without the latent covariate term using the full sample.

4.2 Simulation Study II

4.2.1 Purpose

The second Monte Carlo simulation study: 1) evaluates the ability of the proposed models to provide estimates of the treatment effect for more than just the subpopulation with RV scores at the cutoff, see Section 2.3.2, and 2) compares the performance of the

proposed model to the conventional model in terms of recovering the LATE. Using the proposed model, I quantified the generalizability of the LATE, $ATE_{c\pm1}$, by calculating the ATE within a ±1 point interval around the cutoff $\hat{\theta}_{r,i} = c$ (see Equation 2.10). I quantified the heterogeneity of the LATE by calculating the range of the LATE for the middle 90% of participants, $[LATE_{q.05}, LATE_{q.95}]$, at the cutoff using Equation 2.9. I calculated the ATE within the bottom 30% of participants in the population of the latent construct, ATE_p, using Equation 2.11. In addition, I calculated the LATE with respect to the ORV at the cutoff, $LATE_o$ and the LATE with respect to the LRV at the cutoff, $LATE_l$. Using the conventional model, I calculated the LATE with respect to the latent construct, $LATE_c$. Standard errors for these treatment effect estimates were calculated via application of the delta method (see e.g., Van der Vaart & Wellner, 1996).

4.2.2 Methods

Varying conditions For both the HRD and MRD models, the ICC was fixed at 0.2 as that value is within the range most commonly observed ICCs for cognitive or affective domain variables in education. A pilot study was conducted to determine the sample sizes for both models. Results indicated very similar results when using 500 and 1000 clusters and a poor convergence rate when using 100 clusters (Schochet (2008) suggests at least 100 level-2 units when assignment occurs at the cluster level). Consequently, the number of clusters in this simulation study was varied at 200 and 500. The number of units per cluster will be fixed at 20 for the HRD design; varying this factor is not of interest as the treatment assignment occurs at the cluster level. This results in samples of size 4,000 and 10,000 for the HRD design. In the MRD design the number of units per cluster will be varied at 20 and 40 to represent small and large-sized classrooms, resulting in sample sizes ranging from 4,000 to 80,000. The test length for each latent variable will be varied at 10

Table 4.9: Manipulated Factors in Simulation II.

Manipulated factors	Levels
Test length	(10 items, 30 items)
Level-1 units per cluster	HRD: (20) MRD: (20, 40)
Level-2 units	(200, 500)
Size of LATE	(0, 0.2)
Bandwidth	(1.0, full)

and 30 items using the 2PL model to reflect a short and long survey items in education. The marginal reliability was approximately 0.72 and the Cronbach's alpha was 0.67 for the 10-item condition, which is considered minimally satisfactory. The marginal reliability was approximately 0.89 and the Cronbach's alpha was 0.87 for the 30-item condition which is considered high by education researchers. The cut-point, non-LATE structural parameters, and bandwidth will be the same as those used in simulation study I. The bandwidth will be calculated for the proposed models as in simulation study I. The values of the LATE will be varied at 0.25, a moderate effect size, and 0 to examine Type I error rates. There are a total of 36 conditions that will be tested for the HRD model and 72 conditions that will be tested for the MRD model (see Table 4.9).

Models For simulation study II, two models are fit to the generated data: the proposed model and the conventional RD analysis using summed scores for the latent constructs. As in simulation study I, standard errors for the proposed models were calculated via post-convergence application of the Louis formula (see Section 2.5.3).

Evaluation The generated data were fit to the proposed models used to generate the data as well as to the conventional RD model (Equation3.5). 500 replications were done for each condition. For the proposed model, the $ATE_{c\pm1}$, $[LATE_{q.05}, LATE_{q.95}]$, ATE_p, $LATE_l$, and $LATE_o$ were evaluated in terms of their absolute and relative bias as well as RMSE. Similarly, for the conventional model, $LATE_l$ and $LATE_o$ were evaluated in

terms of absolute and relative bias and RMSE. Under all models, standard errors for these treatment effects were calculated via application of the delta method (see e.g., Van der Vaart & Wellner, 1996) and compared against the Monte Carlo standard deviations.

4.2.3 Results

In this stimulation study, the ability of the HRD and MRD models to recover the average treatment effect (ATE) for participants within one point of the cutoff value, $ATE_{c\pm1}$, the heterogeneity in the local average treatment effect (LATE) due to measurement error in the ORV, $[LATE_{q.05}, LATE_{q.95}]$, the ATE for the bottom 30% of the population, ATE_p, the LATE with respect to the LRV, $LATE_l$, and the LATE with respect to the ORV, $LATE_o$, were evaluated in terms of their bias, relative bias, and RMSE. Similarly, for the conventional model, $LATE_l$ and $LATE_o$ were evaluated in terms of absolute and relative bias and RMSE. Type I error rates for the $LATE_l$ and $LATE_o$ were also evaluated. The HRD models had runtimes between 3 and 14 minutes, while the MRD models had runtimes between 4 and 42 minutes (see Table A.7 in the Appendix). The convergence rates and range of condition numbers of the information matrix for all models fit under simulation study II appear in Table 4.10. Both models have high convergence rates when using the full sample. When using a bandwidth, the convergence rates are lower for the MRD model than the HRD model. Under both models, the convergence rates are better with 30 items than with 10 items when using a bandwidth. Under the poorest performing conditions, less than 75% of MRD models converged when using a 10-item test. As more replications were not run to reach 500 converged analyses, the results discussed in this section are not based on the same number of iterations across conditions. To assess whether the algorithm converged at a local maximum, I checked whether the gradient was sufficiently close to 0. The results appear in Table A.8 in the Appendix.

Table 4.10: Convergence Rates and Condition Numbers

HRD Model								
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Converge	Condition No.	Converge	Condition No.	Converge	Condition No.	Converge	Condition No.
20	200	99%	128 - 147	100%	175 - 217	89%	984 - 1432	91%	651 - 1189
	500	100%	130 - 143	100%	178 - 216	96%	889 - 1402	97%	653 - 1204
MRD Model									
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Converge	Condition No.	Converge	Condition No.	Converge	Condition No.	Converge	Condition No.
20	200	97%	137 - 230	99%	186 - 496	63%	858 - 1699	74%	723 - 1436
	500	98%	126 - 227	99%	164 - 402	72%	792 - 1504	76%	752 - 1448
40	200	100%	158 - 190	100%	190 - 245	86%	744 - 1501	88%	667 - 1234
	500	100%	136 - 167	100%	171 - 230	89%	726 - 1482	93%	634 - 1299

Note: *np* is number of individuals per cluster; Cluster is the number of clusters; Converge is the convergence rate; Condition No. is the condition number of the information matrix

Table 4.11: Bias, Relative Bias, and RMSE for Treatment Effect Estimates under the HRD model with the full sample and bandwidth for 10-item

Clusters	Par	Full Sample			Bandwidth		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
200	$LATE_l$	-0.01	-3%	0.13	0.18	92%	0.24
	$LATE_o$	-0.02	-10%	0.13	0.16	68%	0.23
	$ATE_{c\pm1}$	-0.03	-12%	0.14	0.15	64%	0.23
	$LATE_{q.05}$	-0.05	-7%	0.07	0.25	54%	0.25
	$LATE_{q.95}$	-0.03	-12%	0.06	0.21	46%	0.24
	ATE_p	-0.02	-11%	0.14	0.22	163%	0.29
500	$LATE_l$	<0.01	1%	0.08	-0.06	-29%	0.10
	$LATE_o$	-0.02	-8%	0.08	-0.08	-33%	0.12
	$ATE_{c\pm1}$	-0.02	-8%	0.02	-0.07	-30%	0.11
	$LATE_{q.05}$	-0.01	-1%	0.06	-0.05	-7%	0.06
	$LATE_{q.95}$	-0.06	-11%	0.04	-0.03	-11%	0.04
	ATE_p	<.01	1%	0.09	-0.06	-44%	0.13

HRD Model using the Full Sample and a Bandwidth

Results of the simulation study show that the treatment effect estimates were generally well recovered when using the full sample and tended to be biased when using a bandwidth, especially when the latent variable test length was 10 items. Table 4.11 shows the bias, relative bias, and RMSE for the treatment effect estimates under the HRD model when using 10-item tests for the latent variables. While the bias in the treatment effect estimates is generally less than 0.05 when using the full sample, when conducting the analysis on a subsample of the data defined by a bandwidth of 1 standard deviation of the ORV, the estimates are underestimated (bias between 0.03 and 0.08) with 500 clusters and an overestimated (bias between 0.15 and 0.25) with 200 clusters. Similarly, RMSE values tend to be smaller when using the full data as compared to using a bandwidth and smaller with a sample size of 500 clusters as compared to 200 clusters. RMSE values tend to be smallest for the [$LATE_{q.05}$, $LATE_{q.95}$] estimates and largest for the ATE_p parameter.

When using 30-item tests for the latent variables, treatment effect estimates were well

Table 4.12: Bias, Relative Bias, and RMSE for Treatment Effect Estimates under the HRD model with the full sample and bandwidth for 30-item

Clusters	Par	Full Sample			Bandwidth		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
200	$LATE_l$	-0.02	-12%	0.09	-0.02	-15%	0.11
	$LATE_o$	-0.01	-6%	0.09	-0.02	-10%	0.11
	$ATE_{c\pm1}$	-0.01	-6%	0.08	-0.02	-11%	0.11
	$LATE_{q.05}$	-0.03	-9%	0.05	-0.03	-9%	0.05
	$LATE_{q.95}$	-0.02	-12%	0.04	-0.02	-13%	0.04
	ATE_p	-0.01	-11%	0.14	-0.01	-8%	0.12
500	$LATE_l$	-0.02	-11%	0.06	-0.03	-18%	0.08
	$LATE_o$	-0.02	-8%	0.06	-0.02	-11%	0.07
	$ATE_{c\pm1}$	-0.01	-5%	0.06	-0.02	-10%	0.08
	$LATE_{q.05}$	-0.03	-8%	0.04	-0.03	-9%	0.04
	$LATE_{q.95}$	-0.02	-12%	0.04	-0.02	-12%	0.03
	ATE_p	-0.01	-10%	0.07	-0.02	-14%	0.09

recovered under the full sample conditions and slightly underestimated when using a bandwidth. Table 4.12 shows the bias, relative bias, and RMSE for the treatment effect estimates under the HRD model when using 30-item tests for the latent variables.The bias in treatment estimates is comparable across sample sizes with bias values smaller than 0.03 across conditions. As with the shorter test length, the RMSE values are smaller for the conditions with the larger sample size and the full sample. Furthermore, RMSE values tend to be smallest for the [$LATE_{q.05}$, $LATE_{q.95}$] parameters and largest for the ATE_p parameter.

Type-I error rate is near the nominal level across conditions using the full sample, but inflated when using a bandwidth (See Figure 4.16). Type I error rates tend to be better under conditions using 30-items as opposed to 10-items and 500 clusters as opposed to 200 clusters.

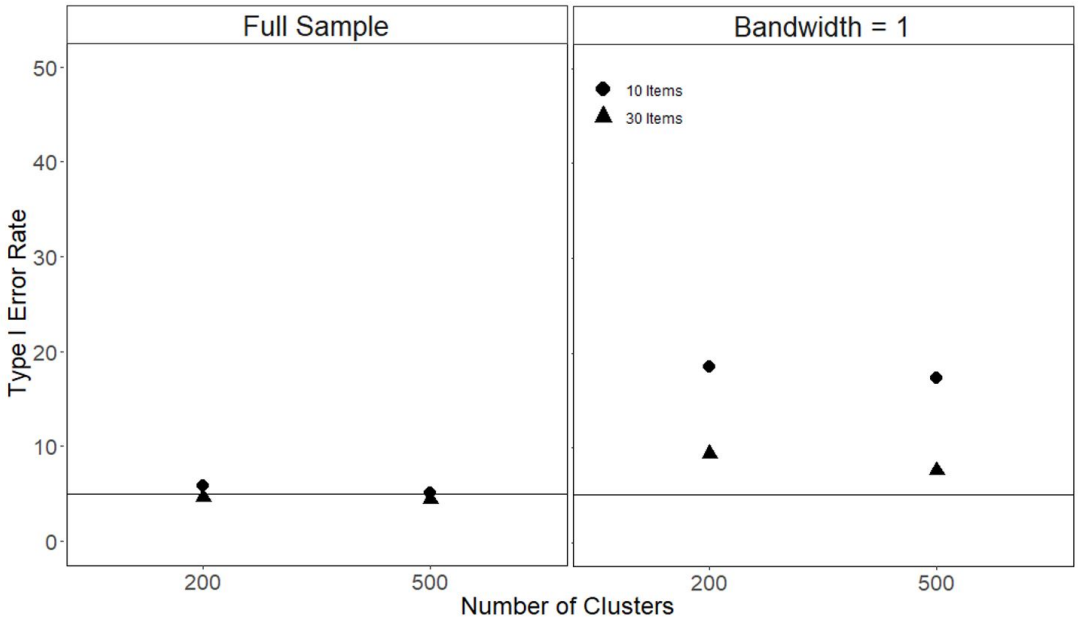

Figure 4.16: Type I error rate of Latent LATE for HRD model using full sample and a bandwidth.

MRD Model using the Full Sample and a Bandwidth

Treatment effect estimates are well recovered under the MRD model when using the full sample and slightly biased when using a bandwidth of 1 standard deviation of the ORV. When using 10-item tests for the latent variables, treatment estimates tend to be underestimated when using a bandwidth (see Table 4.13). The largest attenuation tends to be in the ATE_p parameter. While RMSE values tend to be smaller with a larger sample size, the parameter estimates themselves are largely unaffected by both the number of clusters and the cluster size.

Table 4.13: Bias, Relative Bias, and RMSE for Treatment Effect Estimates under the MRD model with the full sample and bandwidth for 10-item

Clusters	Par	Full Sample						Bandwidth					
		Cluster Size = 20			Cluster Size = 40			Cluster Size = 20			Cluster Size = 40		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
200	$LATE_l$	0.03	12%	0.12	-0.02	-7%	0.09	-0.04	-13%	0.11	-0.02	-7%	0.14
	$LATE_o$	<0.01	2%	0.11	-0.01	-4%	0.09	-0.09	-54%	0.16	-0.07	-38%	0.15
	$ATE_{c\pm1}$	0.02	14%	0.10	-0.01	-5%	0.09	-0.08	-62%	0.19	-0.08	-62%	0.20
	$LATE_{q.05}$	-0.10	-6%	0.05	-0.08	-4%	0.04	-0.14	-7%	0.16	-0.13	-6%	0.14
	$LATE_{q.95}$	-0.01	-3%	0.04	<0.01	2%	0.03	0.01	61%	0.06	0.01	38%	0.04
	ATE_p	0.01	7%	0.10	-0.01	-6%	0.10	-0.10	-75%	0.22	-0.09	-68%	0.20
500	$LATE_l$	<0.01	-1%	0.05	0.02	7%	0.07	-0.04	-18%	0.13	<0.01	2%	0.09
	$LATE_o$	<0.01	-2%	0.06	<0.01	-2%	0.06	-0.02	-11%	0.16	-0.08	-52%	0.15
	$ATE_{c\pm1}$	0.01	4%	0.06	0.02	18%	0.06	-0.04	-28%	0.18	-0.08	-60%	0.16
	$LATE_{q.05}$	-0.11	-7%	0.04	-0.14	-6%	0.02	-0.13	-6%	0.14	-0.13	-6%	0.13
	$LATE_{q.95}$	<0.01	-1%	0.03	<0.01	-1%	0.02	0.01	42%	0.04	0.01	48%	0.03
	ATE_p	<0.01	2%	0.06	0.02	11%	0.06	-0.02	- 14%	0.18	-0.09	-67%	0.17

When using 30-item tests for the latent variables, treatment estimates tend to be overestimated when using a bandwidth (see Table 4.14). The largest bias tends to be in the ATE_p parameter, with the largest bias, 0.15, when using a sample of size $8,000$ with 200 clusters. However, in general, the magnitude of bias in the treatment effect estimates is comparable across overall sample size and number of clusters.

Table 4.14: Bias, Relative Bias, and RMSE for Treatment Effect Estimates under the MRD model with the full sample and bandwidth for 30-item

Clusters	Par	Full Sample						Bandwidth					
		Cluster Size = 20			Cluster Size = 40			Cluster Size = 20			Cluster Size = 40		
		Bias	R. Bias	RMSE	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
200	$LATE_l$	-0.02	-7%	0.13	-0.02	-13%	0.10	0.04	21%	0.16	0.03	15%	0.15
	$LATE_o$	-0.01	-5%	0.13	-0.01	-6%	0.10	0.03	18%	0.18	0.08	51%	0.17
	$ATE_{c\pm1}$	-0.01	-5%	0.14	-0.01	-8%	0.09	0.03	17%	0.17	0.08	52%	0.19
	$LATE_{q.05}$	-0.05	-6%	0.07	-0.05	-3%	0.05	0.07	5%	0.08	0.05	4%	0.07
	$LATE_{q.95}$	-0.01	-5%	0.04	-0.01	-6%	0.03	0.02	8%	0.04	0.01	7%	0.04
	ATE_p	-0.01	-6%	0.12	-0.01	-7%	0.09	0.06	54%	0.25	0.15	114%	0.24
500	$LATE_l$	-0.02	-7%	0.09	-0.02	-11%	0.06	0.05	26%	0.14	0.05	23%	0.09
	$LATE_o$	-0.01	-4%	0.08	-0.01	-4%	0.06	0.08	51%	0.17	0.09	56%	0.12
	$ATE_{c\pm1}$	-0.01	-4%	0.08	-0.01	-8%	0.06	0.08	51%	0.17	0.09	58%	0.13
	$LATE_{q.05}$	-0.06	-5%	0.06	-0.05	-4%	0.03	0.07	6%	0.07	0.05	4%	0.05
	$LATE_{q.95}$	-0.01	-3%	0.03	-0.01	-4%	0.02	0.02	10%	0.03	0.01	5%	0.02
	ATE_p	<0.01	-5%	0.07	-0.01	-6%	0.06	0.07	62%	0.26	0.07	63%	0.20

Type-I error rate is near the nominal level across conditions using the full sample, but inflated when using a bandwidth (See Figure 4.17). Type I error rates tend to be better under conditions using 30-items as opposed to 10-items and 40 individuals per cluster as opposed to 20 individuals per clusters. The Type I error rate does not tend to differ by number of clusters.

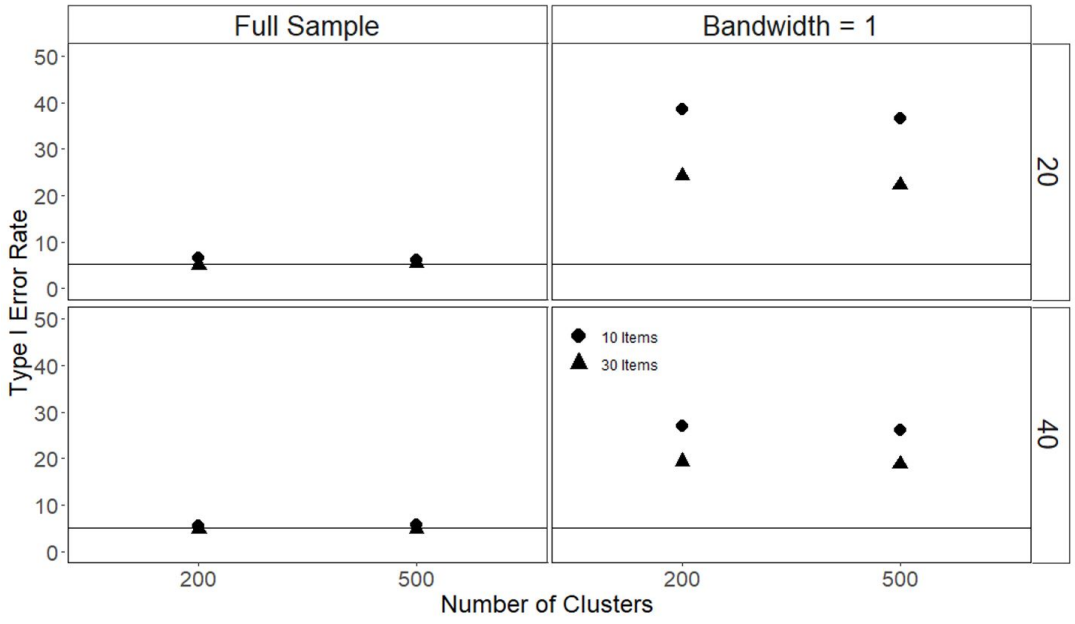

Figure 4.17: Type I error rate of Latent LATE for HRD model using full sample and a bandwidth.

Latent RD Models and Conventional RD Model

The estimation of the LATE with respect to the latent RV is a benefit of the latent RD model over the conventional RD model. As such, I compared the $LATE_l$ from the latent RD models to the LATE with respect to the latent construct from the conventional RD model, $LATE_c$. Figure 4.18 shows the bias in the estimates for the HRD and conventional RD models across sample size and test length. The $LATE_l$ shows bias values smaller than

0.03 for all conditions except when using a bandwidth with 200 clusters and 10 items. The $LATE_c$ estimates are generally more underestimated than the $LATE_l$ across conditions, though the difference in magnitude is small.

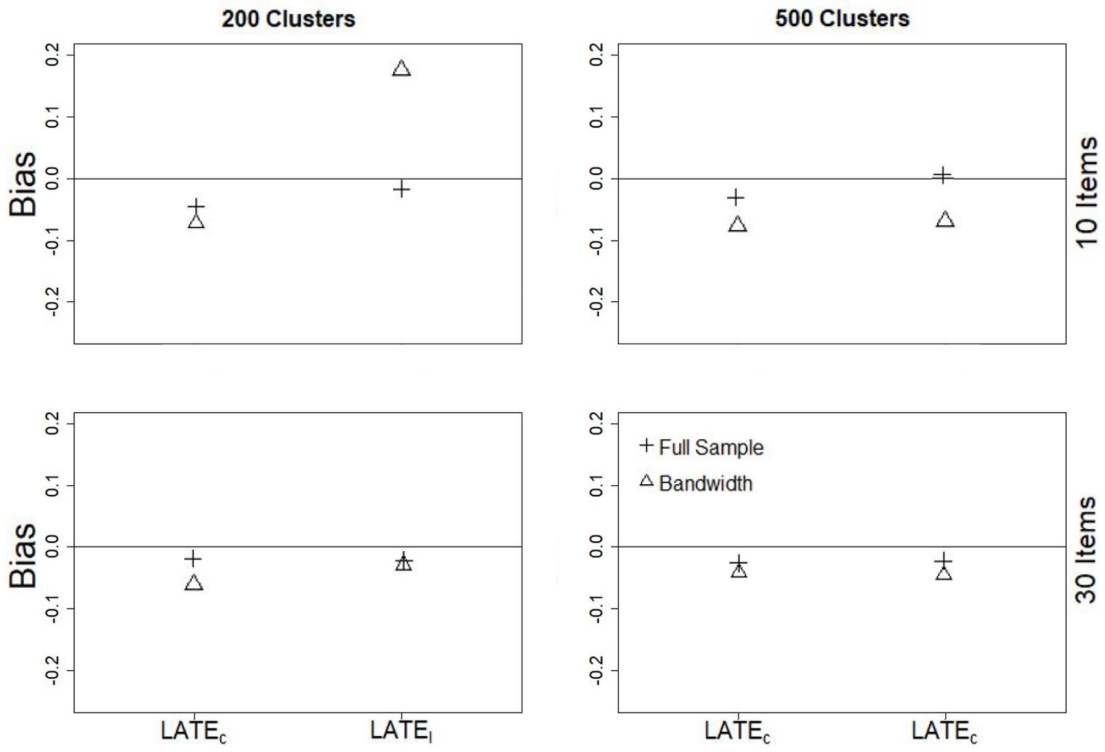

Figure 4.18: Bias in the LATE calculated from the conventional RD model ($LATE_c$) and from the latent HRD model ($LATE_l$) under all simulation conditions.

When using 10-item tests under the MRD model the $LATE_l$ is generally well recovered with bias values no more extreme than 0.05. The $LATE_c$ in the conventional RD model is generally underestimated, especially with the use of a bandwidth. The results are similar when using a 30-item test with the $LATE_l$ being overestimated under most conditions using a bandwidth. Figures 4.19 and 4.20 show the bias in the estimates for the MRD and the conventional RD models across conditions when using 10-item and 30-item tests, respectively.

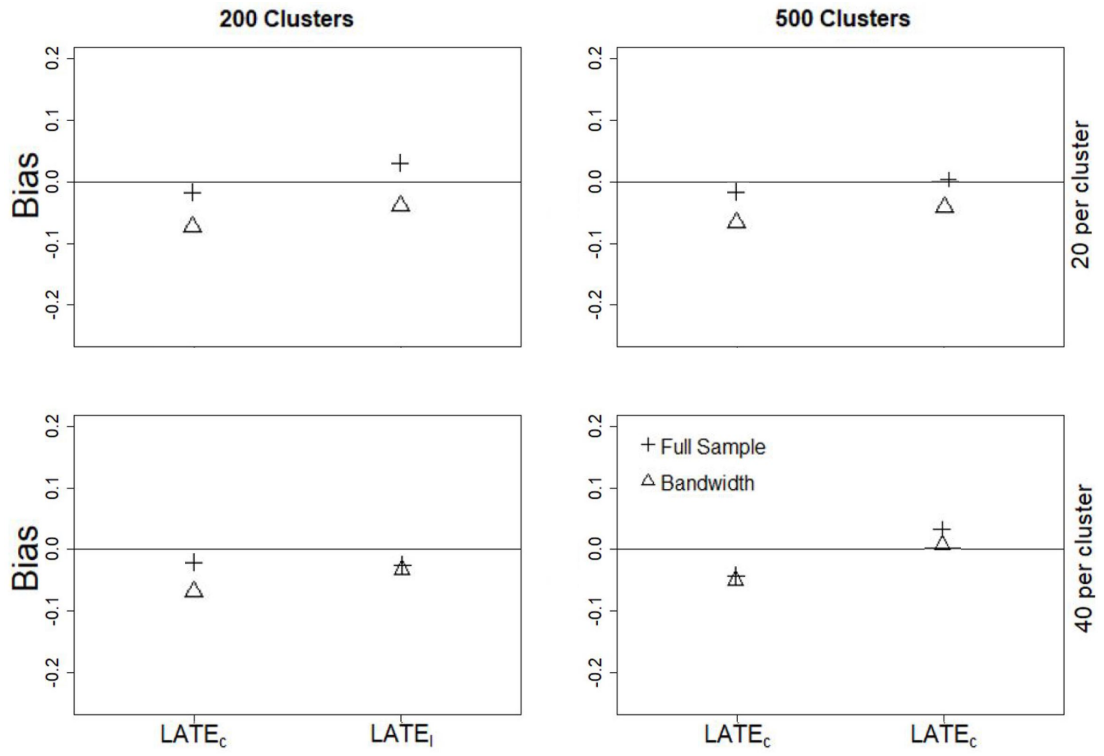

Figure 4.19: Bias in the LATE calculated from the conventional RD model ($LATE_c$) and from the latent MRD model ($LATE_l$) under all simulation conditions with 10-item tests.

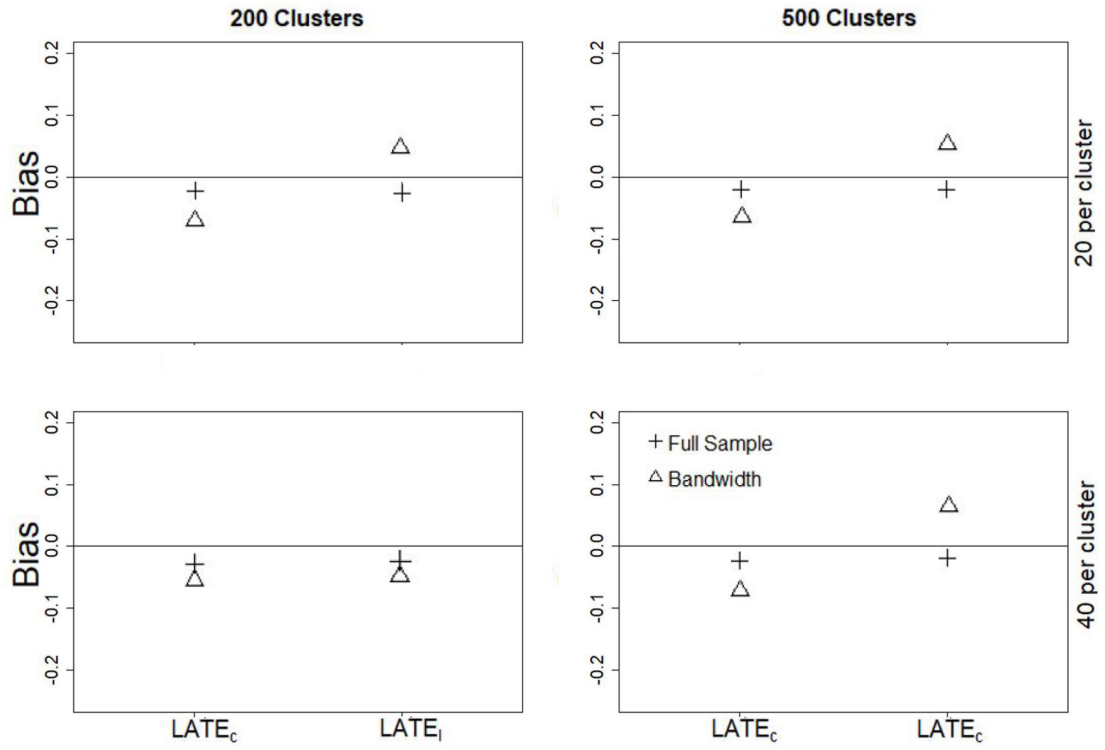

Figure 4.20: Bias in the LATE calculated from the conventional RD model ($LATE_c$) and from the latent MRD model ($LATE_l$) under all simulation conditions with 30-item tests.

Chapter 5: Empirical Data Analysis

In this chapter, I apply the latent variable multisite regression discontinuity model to the Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011 (ECLS-K:2011) data as a mechanical example to illustrate how the proposed model can be used to estimate treatment effects in empirical data. I also apply the conventional RD analysis using observed variables and compare the parameter estimates of both models. This chapter describes the methods and summaries the results of the empirical application.

5.1 Data

The ECLS-K:2011 data is used to examine the effect of being classified as an English language learner (ELL) in kindergarten on students' science achievement in the first grade. Students were given a language screening assessment in the Fall of their kindergarten year. The 20-item assessment consisted of two tasks from the Preschool Language assessment Scale (*pre*LAS; Duncan & De Avila, 2000). Students' ability to follow simple, direct instructions was tested with the "Simon Says" task, and the "Art Show" task tested students' expressive vocabulary. Those who met the minimum score of 17 (i.e., 17 out of 20 items were correctly answered) on the assessment were considered proficient in English; those who did not were classified as ELL. All students were further administered an 18-item measure of cognitive flexibility called the Dimensional Change Card Sort (DCCS; Zelazo, 2006). In the Spring of the first grade all students were administered a 47-item science

assessment (Science).

The sample consists of 18,174 students in 860 schools. Once students were classified based on their *pre*LAS scores, there were 571 schools (n = 15,784) with both students that were designated as ELL and students that were not designated as ELL based on their *pre*LAS scores. While the conventional RD model can estimate the effect of being designated ELL on student's science assessment scores for students with *pre*LAS scores at the cutoff, the multilevel latent RD model can: 1) estimate the effect of being designated ELL on student's science achievement for students with *pre*LAS scores at the cutoff, 2) estimate the treatment effect for students within 1 point of the cutoff, and quantify the heterogeneity in the treatment effect based on measurement error in the *pre*LAS score.

5.2 Analysis

The proposed MRD model was be applied to this data:

$$Science_{ij} = (\gamma_{00} + u_{0j}) + \beta_1 pre\text{LAS}_{ij} + (\gamma_{20} + u_{2j}) ELL_{ij} + \beta_3 pre\text{LAS}_{ij} ELL_{ij} + \beta_4 DCCS_{ij} + \varepsilon_{ij}$$

where $Science_{ij}$ is the latent outcome variable score for individual i in school j based on the first grade science assessment; the variance of the random error term, u_{0j}, is a measure of the between-school variability in science scores; β_1 is the average relation between the running variable, i.e., *pre*LASij, and science achievement; γ_{20} is the average effect of being classified as ELL in kindergarten on science achievement across all schools; $(\gamma_{20} + u_{2j})$ is the LATE for each school; ELL_{ij} is the dichotomous variable that indicates whether a student was designated as ELL, $(ELL_{ij} = 1)$ or as non-ELL $(ELL_{ij} = 0)$; β_4 measures the relation between the DCCS and the science achievement. The LATE at a specific RV is

$\gamma_{20} + \beta_3 \theta_{r,ij}$.

Because only raw number-right scores are provided for the RV, *pre*LAS, and not item-level responses, the item responses were generated in the following way: the number-right values were transformed into scaled summed scores and treated as the true latent variables; the 2PL model was then used to simulate item response data based on the latent variable scores and item parameters generated to be within a reasonable range, i.e., discrimination parameters between 1.0 and 2.5 and difficulties between -2 and 2. The *pre*LAS IRT scores had a marginal reliability of 0.73, which is considered minimally sufficient. Similarly, only IRT factor scores are provided for the outcome variable, *Science*. These scores were used with item parameters, which were generated as previously stated, in the 2PL model to simulate item responses. The simulated item response data were then treated as the real data for the LRV estimation. As item response data is available for the covariate, DCCS, the available item response data was directly used in the MRD model. Missing data on DCCS item responses were treated as incorrect. Students with missing data on *pre*LAS or *Science* were list-wise deleted.

5.3 Results

To tune the MH-RM, I first determined the number of "burn-in" cycles for the MH sampler by examining the auto correlations of random drawings. The time series plots for 4 randomly selected random effects at both levels appear in the Appendix. These plots suggest at least 50 burn-in cycles is reasonable for this model. Next, I ran the full MH-RM algorithm and set the values for the tuning constant w at each level, using the values from Simulation Study II as a guide. The final values were 0.50 at level-1 and 0.04 at level-2. Finally, I ran the full MH-RM algorithm again and examined the trace plots of

the parameter estimates, which appear in the Appendix. The trace plots indicate the initial burn-in stage, $M1$, should be set to 100. The gain constant stage, $M2$ is set to 300 iterations, and the decreasing constant stage, $M3$ is set to 500 iterations.

The structural parameter estimates from the MRD model and the conventional multilevel RD model with observed variables appear in Table 5.1. Under both models there is a significant negative effect of being classified ELL in kindergarten on the outcome. Under the conventional RD model, the treatment effect estimate of -0.27 can only be interpreted as the effect of being classified ELL in kindergarten on students' scores on the science achievement exam in first grade. In contrast, under the MRD model, the treatment effect of -0.12 can be interpreted as the effect of being classified ELL in kindergarten on students' science ability in first grade. This second interpretation allows researchers to make connections with similar studies which have used different science assessments. Furthermore, the MRD model allows for the estimation of the treatment effect for students within 1 point of the cutoff, $ATE_{c\pm1}$, and for a quantification of the heterogeneity in the treatment effect due to measurement error in the ORV, [$LATE_{q.05}$, $LATE_{q.95}$]. The $ATE_{c\pm1}$ is -0.08 (SE = 0.07), indicating that when generalized to students who scored between 16 and 18 on the *pre*LAS, there is no significant effect of being classified ELL on science achievement. The range of the LATE for the middle 90% of students with an ORV of 17 is [0.01, -0.22].

5.4 Limitations

There are several important limitation to this analysis. First, this application is intended as a mechanical example to illustrate the steps and interpretations when applying the proposed MRD model to empirical data. As the full data needed to apply the proposed model was not available, i.e., the individual item responses for the RV (*pre*LAS) and the outcome (Science), this is not a true empirical application. As such, interpretation of the results

Table 5.1: Structural parameter estimates from ECLS-K:2011 data analysis using the MRD model and conventional multilevel RD model

Variable	Latent MRD Model			Conventional Multilevel RD Model		
	$\hat{\beta}$	$SE(\hat{\beta})$	t	$\hat{\beta}$	$SE(\hat{\beta})$	t
*pre*LAS	0.24	0.01	19.18	0.05	0.004	11.41
ELL	-0.12	0.05	-2.43	-0.27	0.05	-4.99
*pre*LAS*ELL*	0.10	0.03	2.91	0.01	0.01	0.69
DCCS	0.32	0.01	31.96	-0.05	0.003	-18.94
$Var(preLAS_j)$	0.01	0.01	2.05	NA	NA	NA
$Var(DCCS_j)$	0.01	0.01	2.08	NA	NA	NA
τ_0	0.01	0.01	2.06	0.09	0.02	4.50
τ_2	0.19	0.02	10.43	0.03	0.03	1.03
τ_{20}	-0.02	0.01	-1.81	-0.02	0.02	-1.09

should only be used for illustrative purposes and should not be taken as a true estimate of the effect of being classified as *ELL* on science achievement.

Second, based on the cutoff value of 17, approximately 86% of the sample was assigned to the treatment condition, while 14% was assigned to the control condition. This is a much more uneven split between treatment conditions than was used in the simulation studies. As such, the effect of this more unbalanced treatment allocation on model parameter estimates is unclear. Results from the simulation study using the MRD model with bandwidths indicates an unequal allocation of participants within clusters may lead to poor model parameter recovery.

Third, the analysis did not take into account background variables such as gender, ethnicity, or socioeconomic status. The treatment effect estimate may change once these covariates are controlled for. Furthermore, the analysis was conducted using the full sample instead of a bandwidth. The linear functional form may not be a proper fit for the data, especially at values away from the cutoff.

Chapter 6: Discussion

6.1 Summary

An important purpose of education research is to examine the causal associations between interventions and student outcomes. The ability of researchers to draw causal conclusions from a study depends on the research design used, as not all designs can support causal inferences. To this end, the regression discontinuity (RD) design is a popular choice in education research as causal effects can be estimated without the need to randomly assign participants to treatment groups. Instead, participants are assigned to treatment groups based on whether their score on a running variable (RV) falls above or below a specified cutoff value. The RD design provides treatment effect estimates for the subpopulation with RV scores at the cutoff, called the local average treatment effect (LATE).

The ability of the RD design to estimate causal effects without requiring the random assignment of participants to treatment groups makes the design a popular choice in education research. Many of the studies conducted in education settings share two features: 1) A lack of independence between individuals in a study due to the hierarchical structure of the education system, and 2) The use of assessments that measure latent constructs. The presence of these characteristics often precludes the use of many traditional statistical models due to violations of the models' assumptions. As such, the use of conventional RD models may not be sufficient to draw valid causal conclusions in many educaitonal research applications. While there has been a trend in many fields towards extending statistical models to

include latent variables (LVs), conventional RD analyses continue to treat latent constructs as observed variables.

This stems from a tradition formalized by Lee and Lemieux (2010), which argues that the measurement error in the observed RV (ORV) converts the RD design into a local RCT at the cutoff and consequently allows researchers to examine the heterogeneity and generalizability of the treatment effect. However, the conventional RD model does not allow for adequate generalizability of the LATE or quantification of the heterogeneity in the treatment effect without the explicit specification of a measurement model for the RV. Furthermore, while any attenuation in the regression parameters due to measurement error in the ORV does not invalidate results when the interpretation of the LATE is confined to the ORV scale, it does make interpretations on the LRV scale invalid. The integration of LV modeling and the RD design would therefore allow researchers to examine the heterogeneity and generalizability of the treatment effect as well as to interpret the LATE with respect to the LRV.

The main purpose of this study was to derive multilevel LV RD models in which the outcome, RV, and covariates are latent and to estimate the proposed models with full-information maximum likelihood (FIML) estimation. Due to the complexity in estimating latent variable models with categorical item responses using traditional FIML algorithms, the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2010a, 2010b) was implemented using R (R Core Team, 2019). Multilevel latent RD models were derived for application when the treatment assignment occurs at the cluster level (hierarchical RD; HRD) and at the individual level (multisite RD; MRD). The performance of the proposed models was examined in two simulation studies.

Results of the first simulation study suggest measurement and structural parameters are properly recovered when analyzing the full sample. This was expected and in line with previous work using a single-level LV RD model and a full sample (Morell et al., 2019).

Standard errors, which were estimated using direct application of the Louis formula, post convergence, were underestimated for all model parameters. However, the 95% confidence interval coverage of the structural parameters and the $LATE_l$ tended to be near the nominal level for both models.

When only a subsample of participants with ORV values within 1 standard deviation of the cutoff is used (i.e., when a bandwidth of 1 is used), several structural parameters were biased. As expected, the measurement parameters were well recovered under conditions using a bandwidth because they were estimated using the full sample. The bias in structural parameters tended to be more severe for the MRD model. However, the variance of the cluster-level RV values, $Var(\theta_{r,j})$, were severely underestimated (up to 85%) under both models. This attenuation is expected as the use of a bandwidth specifically restricts the range of RV values and, therefore, the between cluster variability. The relation between the RV and the outcome variable, β_1, was underestimated by about 50% with the MRD model. Accordingly, the interaction between the RV and the treatment assignment was overestimated by 140%. This overestimation is an interaction between the the assignment mechanism in the MRD model (i.e., participants are assigned at the individual level) and the use of a bandwidth (i.e., each cluster only contains participants with a narrow band of values on the RV). This bias was not seen in the HRD model because, as participants are assigned at the cluster-level, all values of the RV are still represented in the data, within each cluster.

Standard errors tended to be more severely underestimated for the structural parameter when using a bandwidth than when using the full sample. This is due to the use of a "multi-stage" estimation approach when using a bandwidth. The measurement parameters, which were first estimated using the full sample, were treated as fixed when only the data within the bandwidth was fit to the model. Treating estimated parameters as fixed in a subsequent stage of estimation is known to underestimate standard errors (Y. Liu, Yang, &

Maydeu-Olivares, 2019), a point I return to in the next section.

The 95% confidence interval coverage of the structural parameters when using a bandwidth tended to be poor across both model. The regression parameters had especially low coverage under the MRD model. However, the $LATE_l$ had coverage near 90% under both models. As expected, due to the nature of using only the data within a bandwidth, the variance of the cluster-level RV values, $Var(\theta_{r,j})$, had 0% coverage, while the variance of the cluster-level latent covariate values , $Var(\theta_{c,j})$, had about 75% coverage.

In order to assess the impact of a violation of the latent RD model assumption of a properly specified model, two misspecified models were examined under the HRD and the MRD models, one in which the interaction term is omitted and one in which the latent covariate is omitted. The measurement parameters were generally well recovered, with the outcome's item parameters being slightly over estimated under both misspecifications. When the interaction term was omitted, the structural parameters were properly recovered under the HRD model with the largest relative bias, -14%, in the treatment effect parameter, β_2. However, many of the structural parameters under the MRD model were attenuated. The relative bias in γ_{20} was more than twice that seen in the HRD model, -29%. Furthermore, the variance of the level-2 random effects for the LATE , τ_2^2 had a relative bias of -21%. Consequently, $LATE_l$ was properly recovered under the HRD, but attenuated under the MRD. Similarly, when the latent covariate is omitted, the structural parameters are well recovered under the HRD, but the interaction term, β_3 is overestimated with a relative bias of 20%, and the variance of the level-2 random effects for the LATE, τ_2^2 is similarly underestimated. Under both conditions, the assignment mechanism in the MRD model (i.e., assignment at the individual level) appears to be more sensitive to misspecifications than that of the HRD model (i.e., assignment at the cluster level). The results highlight the importance of assessing the fit of the latent RD model before interpreting the results, especially when using the MRD model. Furthermore, as these misspecifications occurred when

simulation conditions were ideal, parameter bias may be worse under some more practical conditions.

A second simulation study was conducted to examine the ability of the proposed models to quantify the heterogeneity and generalizability of the treatment effect and to compare the LATE estimates between the latent multilevel models and the conventional RD model with observed variables. As with the previous simulation study, the results indicate the models perform well when using the full sample. While the use of a bandwidth generally resulted in worse estimates, the magnitude of the bias in parameter estimates depended largely on the cluster size and the test lengths for the LVs. The parameter recovery was worse with a test length of 10 than with a test length of 30. This may reflect the impact of the reliability of the LRV, as the 10-item condition had a marginal reliability of approximately 0.72 and the 30-item conditions had a marginal reliability of approximately 0.89, and measures with lower reliability tend to attenuate parameter estimates. The parameter estimates tended to be comparable under 200 and 500 clusters with the exception of the HRD model using 10-item tests. Under this condition, the smaller cluster size resulted in relative bias values between 46 and 140%. The number of participants within each cluster did not affect parameter estimates under the MRD model. Similarly, the Type I error rate was well controlled when using the full sample and elevated with a bandwidth, with worse performance under the MRD model than the HRD model.

One of the benefits of the latent RD model over the conventional RD model is the ability to interpret the LATE with respect to the LRV as well as the ORV. Using the conventional RD model to interpret the LATE with respect to the LRV can be considered a misspecification of the measurement model, as observed scores are used for the LVs. The recovery of the LATE with respect to the LRV, $LATE_l$ in the latent RD models was compared to that in the conventional RD model, $LATE_c$. The bias was comparable under the HRD and the conventional models across conditions, though the HRD model tended to have less biased

estimates. The $LATE_l$ under the MRD model tended to be less biased than the $LATE_c$ across conditions. While the magnitude in the bias between the MRD and the conventional RD model was greater than that between the HRD and the conventional RD model, the $LATE_c$ parameter estimates tended to exhibit bias less than .07 across conditions. This suggests the LATE with respect to the latent construct may still be estimated accurately when using the conventional RD model when a sufficient number of items is used.

The two simulation studies indicate that the proposed models perform well with large sample sizes, moderate to large tests with adequate reliability, and the full sample. While the use of a bandwidth results in worse estimates across models, the MRD model is more sensitive to the range restriction due to the assignment of participants to treatment groups at the individual level. Moreover, when bandwidths are used with the MRD model, the design becomes unbalanced, which may also contribute to the difficulties in estimation. The results also suggest sample size recommendations for practitioners using the models. When using the full sample to estimate the treatment effect, as few as 200 clusters and 20 individuals per cluster may be sufficient when the LRV is of at least moderate length and has adequate reliability.

A second purpose of this study was to provide an empirical illustration using data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-2011 (ECLS-K:2011). An MRD model was fit to the data to estimate the causal effect of being designated as an English language learner (ELL) in kindergarten on students' science achievement in first grade. The regression parameters were underestimated in the conventional RD model as compared to the latent MRD model. However, both the MRD and the conventional RD model found a significant negative effect. The MRD model was also able to provide more information about the treatment by estimating a treatment effect for students within 1 point of the *ELL* designation cutoff, which indicates that once the effect is generalized to a wider range of scores, the treatment may not have a significant effect on

science achievement. Furthermore, the quantification of the heterogeneity in the treatment effect due to the measurement error in the *pre*LAS scores showed a range from no effect to a stronger negative effect.

The results of the empirical analysis must be interpreted in light of the results of the simulation studies. The MRD model tended to have more biased estimates than the HRD model across conditions. It was also more sensitive to changes in simulation conditions. While the exact conditions of the empirical data were not used in any simulation study, the number of clusters (571), the number of students per cluster (2 to 26), and the LRV with 20 items and a marginal reliability of 0.73 is similar to conditions examined in the simulation studies using a bandwidth and a 10-item test, suggesting that the treatment effect in the empirical analysis may be underestimated. Furthermore, the current study did not apply sampling weights, which is common practice in the use of such large scale data.

6.2 Directions for Future Study

The findings and limitations of this study suggest several areas for future research in expanding the applicability and usefulness of the latent multilevel RD model. First, the standard errors tended to be underestimated across both models. Per large sample theory, standard errors should be unbiased using the delta method. Furthermore, when using restricted maximum likelihood estimation both parameter variances and standard errors should be unbiased. However, the current study uses a finite sample, which may be resulting in biased standard errors. Further investigations are needed to determine if the direct application of the Louis formula is appropriate for these models. For example, the number of iterations used may need to be increased further; however, as this does increase the already long runtime of the models. Additionally, in the current study only one sample is drawn from the MH sampler at a time. Drawing multiple samples at each iteration may

improve the standard error estimates.

Second, when using multilevel modeling it is standard practice to fit a series of models and conduct nested model comparisons. However, the current implementation of the MH-RM algorithm does not calculate the likelihood needed to conduct nested model comparisons or to calculate fit indices such as the Akaike Information Criterion (AIC; Akaike, 1987) or the Bayesian Information Criterion (BIC; Schwarz et al., 1978).

Third, the exploratory fitting of misspecified models highlighted the importance of assessing model fit when using the latent multilevel RD model. As such, goodness of fit statistics should be developed for this purpose (e.g., Bock & Aitkin, 1981; Browne & Cudeck, 1993; Maydeu-Olivares & Joe, 2005, 2006). Furthermore, in order to support generalizing the ATE to ORV values away from the cutoff, a series of models may be fit using varying bandwidths. Adequate model fit may be used to justify generalizing the treatment effect to ORV values based on the bandwidths used.

Fourth, the use of bandwidths with the latent multilevel RD model is more complicated than with the conventional RD model. Using only the data that falls within the specified bandwidth to estimate the full latent RD model would mean the measurement model for the RV is misspecified as the response data within the bandwidth does not represent the full distribution of ability values. To avoid this issue, the approach used in this study was to estimate the item parameters first using the full data and then treat them as fixed and estimate the structural parameters using the bandwidth data. However, this approach has its own disadvantages as the same item response data should not be used to estimate item parameters (i.e., calibration) and to calculate ability estimates (i.e., scoring). Furthermore, treating the item parameter estimates as fixed in the second stage of the estimation results in underestimated standard errors, which can be corrected using the restricted recalibration approach (Y. Liu, Yang, & Maydeu-Olivares, 2019).

Fifth, the performance of the proposed models should be examined under a greater

variety of sample sizes, test lengths, and number of random effects to better understand the relationship between these factors and its performance. When using a single-level LV RD model, Morell et al. (2019) found the model recovered parameters well even with a test-length of 5 and sample size of 500. Finding the limitations of the proposed models in terms of sample size and test length will be beneficial. Similarly, the MRD model used specified no relationship between the random intercept and the random slope terms. It is also possible that there is a relationship between these two terms which may affect the recover of structural parameters.

Sixth, while the MRD model allows for cutoff values to vary across clusters, the extent to which these values may vary is unclear. The interpretation of the overall treatment effect when there is little or no overlap in ORV values across clusters must be done carefully. Furthermore, the effect of a large difference in the proportion of individuals assigned to each treatment condition across clusters need to be further investigated, as the results of the current study showed poor parameter recovery when a bandwidth waws used with the MRD model.

Seventh, the proposed models are classified as "sharp" RD designs as all participants comply with their treatment assignment; however, in practice, it may be more likely that the not all participants adhere to their treatment assignment, i.e., the RD design is fuzzy. The proposed models may still be applied in this scenario by including the actual treatment indicator instrumented by the eligibility indicator.

Eighth, the proposed model may be extended to accommodate multi-dimensional measurement structures, such as a two-correlated factor model, which are common in empirical data. The added complexity of such measurement models may complicate the estimation of the multilevel LV RD models.

Appendix A: Appendix

Table A.1: Bias, Relative Bias, and RMSE for Running Variable Slopes for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
1.41	0.02	1%	0.05	0.01	1%	0.04
1.00	0.01	1%	0.03	0.01	1%	0.03
1.07	0.01	1%	0.02	0.01	1%	0.03
1.61	0.03	2%	0.05	0.03	2%	0.05
1.23	0.01	1%	0.04	0.01	1%	0.04
1.16	0.01	1%	0.04	0.01	1%	0.04
1.35	0.02	1%	0.05	0.01	1%	0.04
1.24	0.01	1%	0.04	0.02	1%	0.04
1.38	0.02	1%	0.05	0.02	1%	0.05
1.63	0.02	1%	0.05	0.01	1%	0.04
1.26	0.01	1%	0.05	0.01	1%	0.05
1.76	0.02	1%	0.05	0.01	1%	0.05
1.65	0.02	1%	0.05	0.01	1%	0.05
1.31	0.01	1%	0.04	0.01	1%	0.04
1.20	0.01	1%	0.04	0.02	1%	0.04
1.43	0.02	1%	0.05	0.01	1%	0.04
1.30	0.02	1%	0.04	0.01	1%	0.04
1.29	0.01	1%	0.04	0.01	1%	0.04
1.40	0.01	1%	0.05	0.02	1%	0.05
1.36	0.02	1%	0.04	0.02	1%	0.04
1.71	0.03	1%	0.06	0.01	1%	0.05
1.21	0.01	1%	0.04	0.01	1%	0.04
1.42	0.02	1%	0.04	0.01	1%	0.04
1.66	0.01	1%	0.05	0.01	1%	0.05
1.45	0.02	1%	0.04	0.01	1%	0.04
1.34	0.01	1%	0.04	0.01	1%	0.04
1.32	0.01	1%	0.04	0.01	1%	0.04
1.26	0.01	1%	0.04	0.02	1%	0.03
1.11	0.02	1%	0.03	0.01	1%	0.03
1.48	0.02	1%	0.05	0.01	1%	0.04

Table A.2: Bias, Relative Bias, and RMSE for Running Variable Intercepts for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
-0.35	-0.03	9%	0.07	-0.02	8%	0.06
0.68	-0.01	-1%	0.05	-0.01	-1%	0.05
-0.31	-0.02	6%	0.05	-0.02	6%	0.05
-1.96	-0.03	2%	0.09	-0.02	2%	0.09
-1.03	-0.02	2%	0.06	-0.03	3%	0.07
-1.56	-0.02	1%	0.06	-0.02	1%	0.06
0.73	-0.02	-3%	0.07	-0.02	-3%	0.07
-1.91	-0.02	1%	0.07	-0.02	1%	0.07
0.64	-0.02	-3%	0.07	-0.01	-2%	0.07
-1.12	-0.03	3%	0.07	-0.02	3%	0.07
1.99	-0.02	-1%	0.09	-0.02	-1%	0.09
1.18	-0.03	-3%	0.07	-0.02	-3%	0.07
0.44	-0.02	-5%	0.09	-0.02	-5%	0.09
-2.42	-0.02	1%	0.09	-0.02	1%	0.09
1.06	-0.02	-2%	0.08	-0.02	-2%	0.08
0.84	-0.02	-2%	0.07	-0.02	-2%	0.07
-1.70	-0.02	1%	0.08	-0.02	1%	0.08
1.14	-0.02	-2%	0.07	-0.02	-2%	0.07
2.33	-0.02	-1%	0.07	-0.03	-2%	0.07
0.96	-0.02	-2%	0.08	-0.01	-1%	0.08
-0.64	-0.03	5%	0.07	-0.02	4%	0.07
-0.96	-0.01	1%	0.09	-0.01	1%	0.09
1.21	-0.02	-2%	0.07	-0.02	-2%	0.07
1.91	-0.02	-1%	0.08	-0.02	-1%	0.07
0.31	-0.02	-6%	0.09	-0.02	-6%	0.08
-0.67	-0.02	3%	0.08	-0.02	3%	0.08
2.09	-0.02	-1%	0.07	-0.03	-1%	0.07
0.47	-0.02	-4%	0.07	-0.01	-4%	0.07
-0.56	-0.01	2%	0.06	-0.02	3%	0.07
0.54	-0.03	-6%	0.08	-0.03	-6%	0.08

Table A.3: Bias, Relative Bias, and RMSE for Covariate Slopes for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
1.70	0.01	1%	0.05	0.01	1%	0.05
1.55	0.01	<1%	0.05	0.01	<1%	0.05
1.31	0.01	<1%	0.04	0.01	<1%	0.04
1.35	0.01	<1%	0.04	0.01	1%	0.03
1.41	0.01	1%	0.04	0.01	1%	0.05
1.19	0.01	<1%	0.04	0.01	<1%	0.04
1.28	0.01	<1%	0.03	0.01	<1%	0.03
1.46	0.01	1%	0.04	0.01	1%	0.04
1.57	0.01	1%	0.04	0.01	1%	0.04
1.13	0.01	1%	0.03	0.01	<1%	0.03
1.18	0.01	<1%	0.04	0.01	<1%	0.04
1.09	0.01	<1%	0.04	0.01	<1%	0.04
1.86	0.01	<1%	0.05	0.01	1%	0.06
1.35	0.01	1%	0.04	0.01	1%	0.04
1.68	0.01	1%	0.04	0.01	<1%	0.04
1.41	0.01	<1%	0.05	0.01	<1%	0.05
1.50	0.01	1%	0.05	0.01	1%	0.05
1.10	0.01	<1%	0.03	0.01	1%	0.03
1.48	0.01	<1%	0.04	0.01	<1%	0.03
1.41	0.01	1%	0.04	0.01	<1%	0.03
1.22	0.01	<1%	0.03	0.01	<1%	0.03
1.21	0.01	1%	0.04	0.01	1%	0.04
1.45	0.01	1%	0.04	0.01	1%	0.04
1.37	0.01	1%	0.04	0.01	1%	0.04
1.58	0.01	1%	0.04	0.01	1%	0.04
1.28	0.01	<1%	0.03	0.01	1%	0.04
1.08	0.01	<1%	0.03	0.01	<1%	0.03
1.10	0.01	1%	0.03	0.01	<1%	0.03
1.84	0.01	1%	0.05	0.01	1%	0.04
1.26	0.01	1%	0.04	0.01	1%	0.04

Table A.4: Bias, Relative Bias, and RMSE for Covariate Intercepts for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
0.57	0.02	3%	.09	0.02	3%	.09
2.67	0.02	1%	.09	0.02	1%	.09
-1.67	0.02	-1%	.07	0.02	-1%	.07
0.64	0.01	2%	.07	0.01	2%	.07
-0.27	0.01	-4%	.06	0.01	-3%	.05
-1.35	0.01	-1%	.07	0.01	-1%	.07
-0.48	0.01	-2%	.07	0.01	-2%	.07
-0.31	0.02	-6%	.06	0.01	-3%	.05
-0.40	0.02	-5%	.09	0.01	-3%	.08
-0.88	0.01	-1%	.06	0.02	-2%	.07
1.41	0.01	1%	.07	0.02	1%	.07
-1.40	0.01	-1%	.06	0.02	-1%	.06
0.69	0.02	3%	.09	0.01	1%	.07
1.21	0.02	2%	.07	0.02	2%	.07
-0.38	0.02	-5%	.09	0.01	-3%	.08
2.54	0.02	1%	.09	0.03	1%	.09
-2.87	0.02	-1%	.09	0.02	-1%	.09
-1.01	0.01	-1%	.06	0.02	-2%	.07
-0.85	0.02	-2%	.08	0.02	-2%	.08
-0.28	0.02	-7%	.08	0.01	-4%	.06
0.14	0.01	7%	.06	0.01	7%	.06
0.29	0.01	3%	.07	0.02	7%	.09
0.51	0.02	4%	.08	0.02	4%	.08
-0.37	0.02	-5%	.07	0.01	-3%	.06
-0.11	0.02	-19%	.09	0.01	-9%	.05
0.13	0.01	8%	.08	0.02	16%	.10
-1.00	0.01	-1%	.06	0.03	-3%	.07
1.03	0.01	1%	.06	0.02	2%	.06
0.46	0.02	4%	.09	0.01	2%	.08
-2.36	0.02	-1%	.07	0.02	-1%	.06

Table A.5: Bias, Relative Bias, and RMSE for Outcome Variable Slopes for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
1.41	-.03	-2%	.04	-.03	-2%	.04
1.48	-.03	-2%	.05	-.02	-2%	.04
1.01	-.02	-2%	.03	-.02	-2%	.03
1.42	-.03	-2%	.05	-.03	-2%	.05
1.25	-.02	-2%	.04	-.02	-2%	.04
1.22	-.02	-2%	.04	-.02	-2%	.04
1.46	-.03	-2%	.05	-.03	-2%	.05
1.08	-.02	-2%	.03	-.01	-1%	.02
1.50	-.03	-2%	.05	-.03	-2%	.05
1.25	-.02	-2%	.04	-.02	-2%	.04
1.47	-.03	-2%	.05	-.02	-2%	.05
1.50	-.03	-2%	.05	-.03	-2%	.05
1.22	-.02	-2%	.04	-.03	-2%	.04
1.19	-.02	-2%	.04	-.03	-2%	.04
1.35	-.03	-2%	.04	-.03	-2%	.04
1.19	-.02	-2%	.04	-.02	-2%	.04
1.15	-.02	-2%	.03	-.02	-2%	.03
1.37	-.02	-2%	.05	-.01	-1%	.04
1.41	-.03	-2%	.05	-.03	-2%	.05
1.38	-.02	-2%	.05	-.02	-2%	.05
1.24	-.02	-2%	.04	-.01	-1%	.04
1.57	-.02	-2%	.05	-.02	-2%	.05
1.42	-.02	-2%	.04	-.02	-2%	.04
1.48	-.02	-2%	.04	-.03	-2%	.04
1.51	-.03	-2%	.05	-.03	-2%	.05
1.27	-.02	-2%	.04	-.02	-2%	.04
1.48	-.03	-2%	.04	-.03	-2%	.04
1.19	-.02	-2%	.04	-.03	-3%	.04
1.09	-.02	-2%	.03	-.02	-2%	.03
1.57	-.03	-2%	.05	-.03	-2%	.05

Table A.6: Bias, Relative Bias, and RMSE for Outcome Variable Intercepts for HRD and MRD Models in Simulation Study I using Full Sample

True		HRD			MRD	
Value	Bias	R. Bias	RMSE	Bias	R. Bias	RMSE
-0.55	-0.11	20%	0.11	-0.11	20%	0.11
1.29	-0.12	-9%	0.12	-0.11	-9%	0.08
1.38	-0.05	-4%	0.05	-0.06	-4%	0.05
-1.21	-0.11	9%	0.11	-0.11	9%	0.11
-0.63	-0.08	13%	0.09	-0.09	14%	0.09
2.22	-0.08	-4%	0.09	-0.07	-3%	0.09
-2.13	-0.11	5%	0.12	-0.10	5%	0.11
-1.43	-0.06	4%	0.06	-0.06	4%	0.06
0.41	-0.12	-30%	0.13	-0.13	-32%	0.14
1.21	-0.08	-7%	0.09	-0.10	-8%	0.09
-0.40	-0.12	30%	0.12	-0.11	27%	0.11
-0.44	-0.12	28%	0.12	-0.12	27%	0.12
-0.28	-0.08	29%	0.08	-0.09	33%	0.10
2.00	-0.07	-4%	0.08	-0.08	-4%	0.08
0.55	-0.10	18%	0.10	-0.11	-20%	0.10
-.60	-0.08	13%	0.08	-0.09	15%	0.09
0.01	-0.07	-2549%	0.07	-0.07	-2534%	0.07
1.69	-0.10	-6%	0.11	-0.10	-1%	0.05
0.33	-0.11	-33%	0.11	-0.11	-3%	0.06
-2.25	-0.10	5%	0.11	-0.10	4%	0.10
0.49	-0.08	-17%	0.09	-0.10	-20%	0.10
2.00	-0.13	-7%	0.14	-0.11	-6%	0.14
1.15	-0.11	-10%	0.11	-0.11	-10%	0.11
0.07	-0.12	-171%	0.12	-0.12	-173%	0.12
1.23	-0.12	-10%	0.13	-0.12	-10%	0.13
-1.88	-0.09	5%	0.09	-0.10	5%	0.09
0.03	-0.12	-380%	0.12	-0.11	-353%	0.11
-0.08	-0.08	90%	0.08	-0.08	95%	0.09
0.52	-0.06	-12%	0.07	-0.06	-12%	0.07
0.47	-0.13	-29%	0.14	-0.13	-28%	0.14

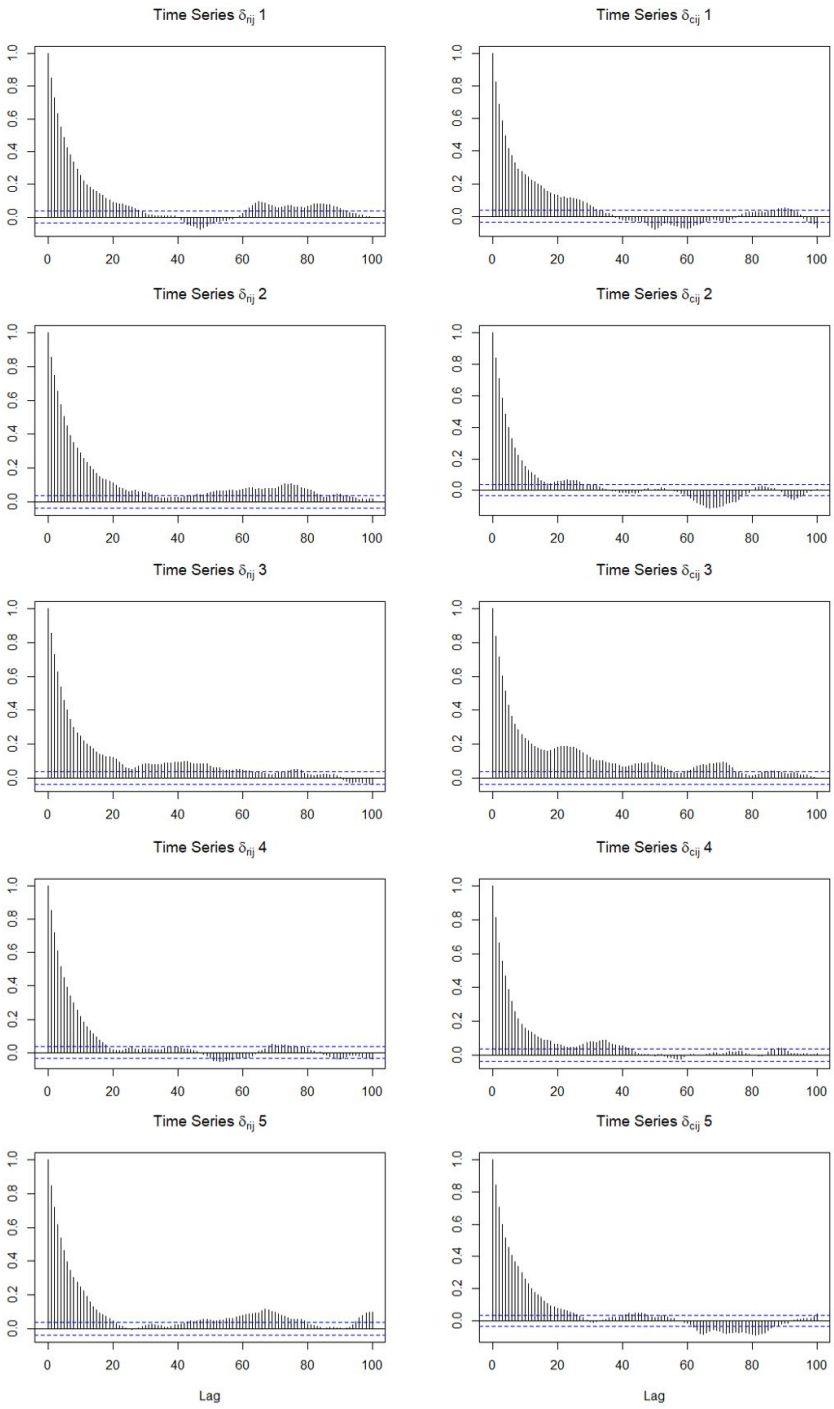

Figure A.1: Time-series plots for 5 randomly chosen δ_{rij} andδ_{cij} under the HRD model, sample size = 4,000, number of clusters = 200

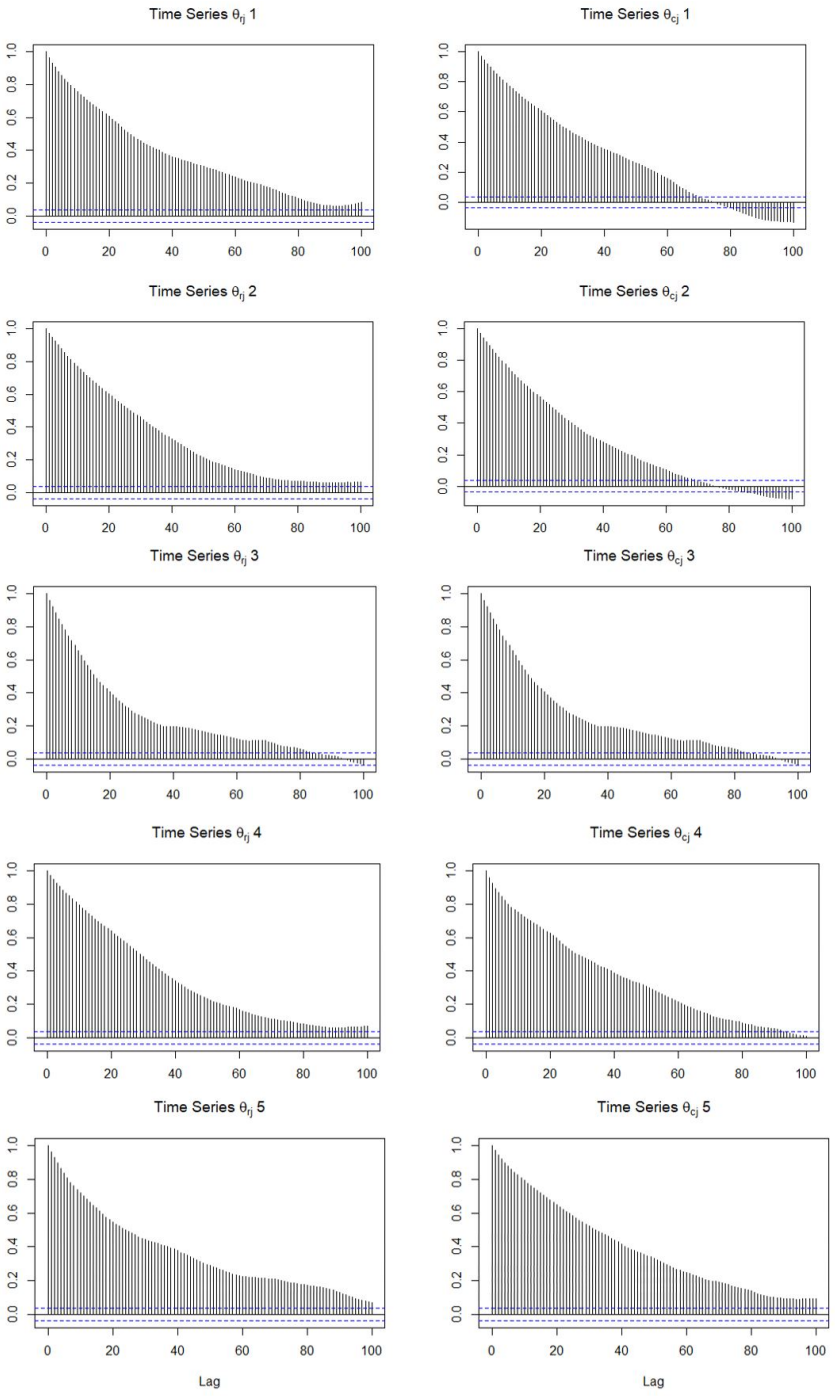

Figure A.2: Time-series plots for 5 randomly chosen θ_{rj} andθ_{cj} under the HRD model, sample size = 4,000, number of clusters = 200

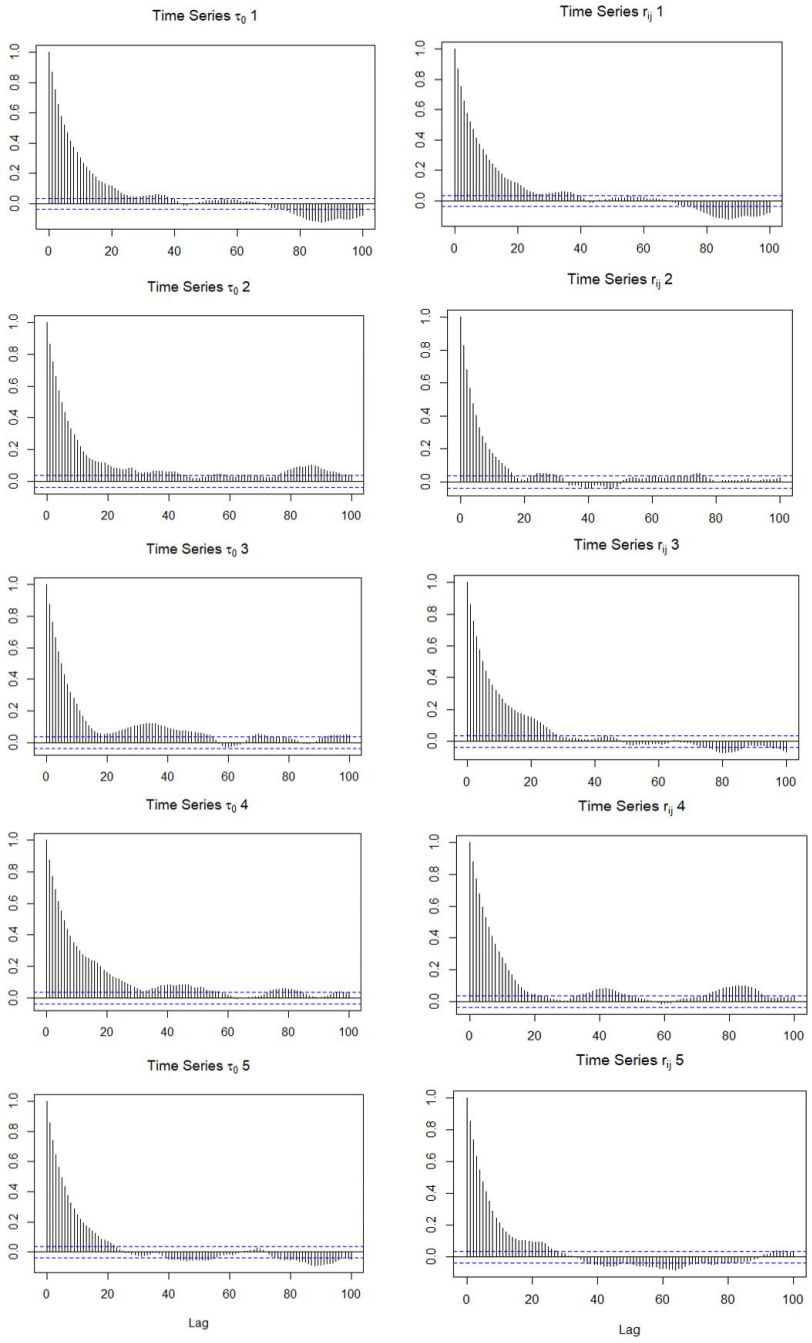

Figure A.3: Time-series plots for 5 randomly chosen τ_{0j} andε_{ij} under the HRD model, sample size = 4,000, number of clusters = 200

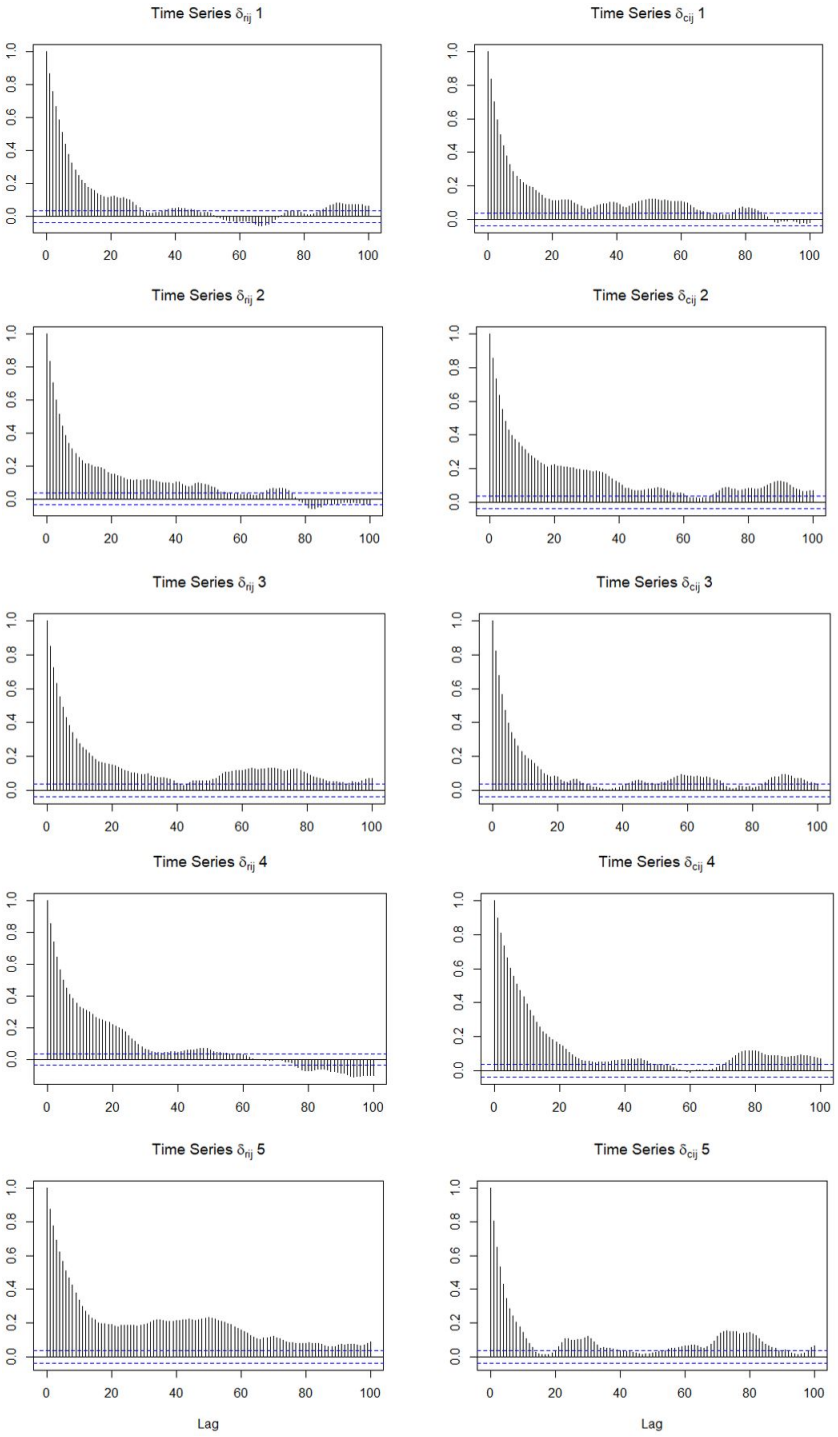

Figure A.4: Time-series plots for 5 randomly chosen δ_{rij} andδ_{cij} under the MRD model, sample size = 4,000, number of clusters = 200

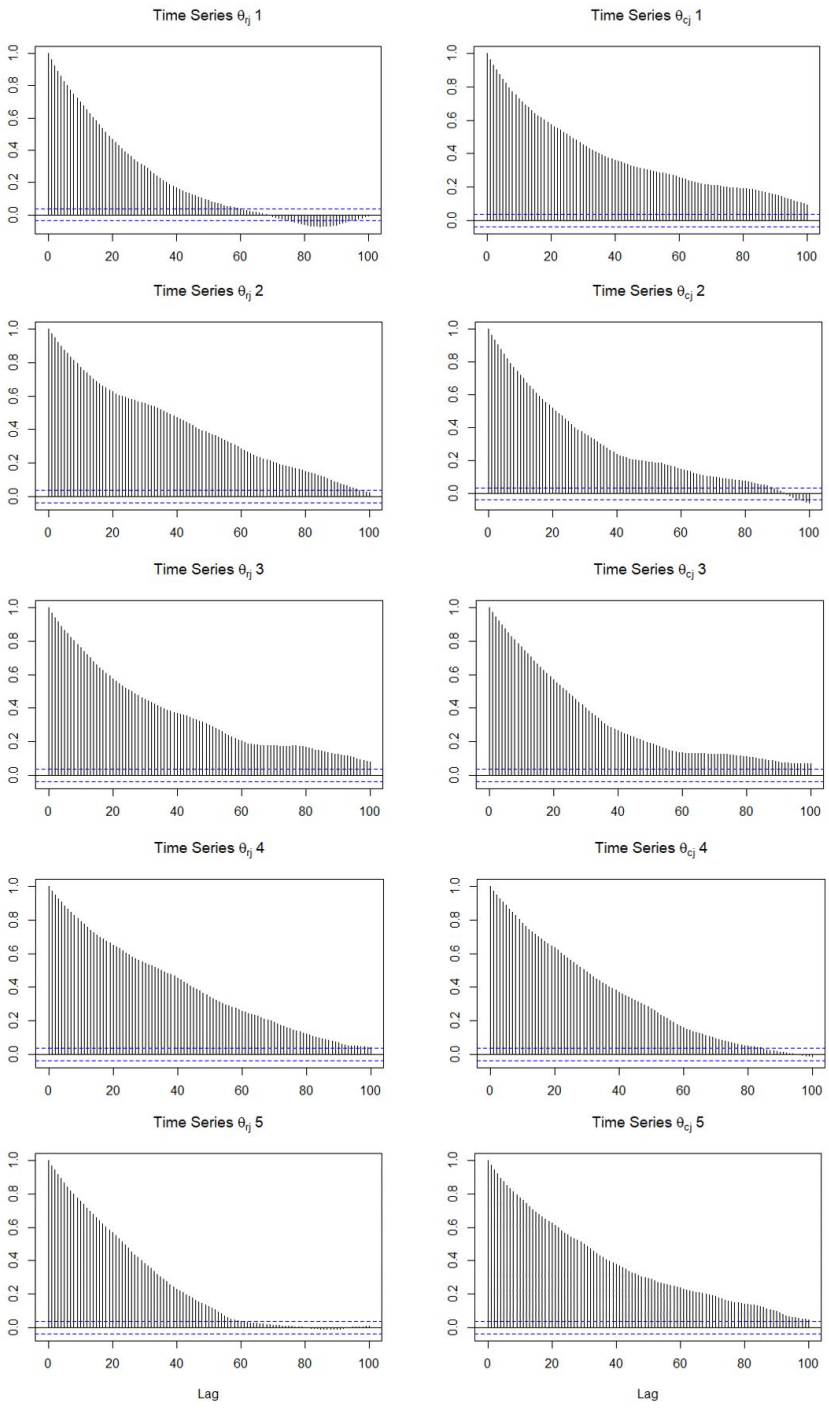

Figure A.5: Time-series plots for 5 randomly chosen θ_{rj} andθ_{cj} under the MRD model, sample size = 4,000, number of clusters = 200

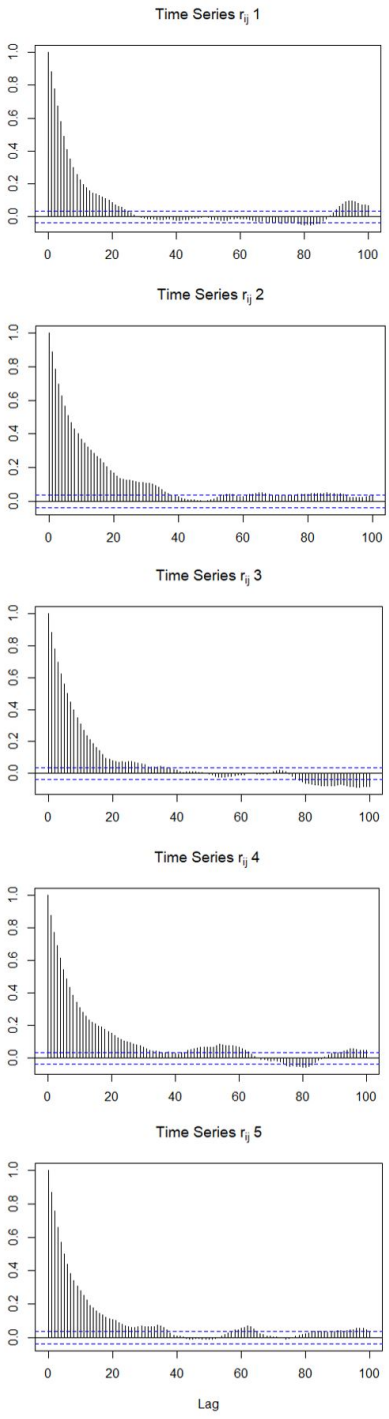

Figure A.6: Time-series plots for 5 randomly chosen ε_{ij} under the MRD model, sample size = 4,000, number of clusters = 200

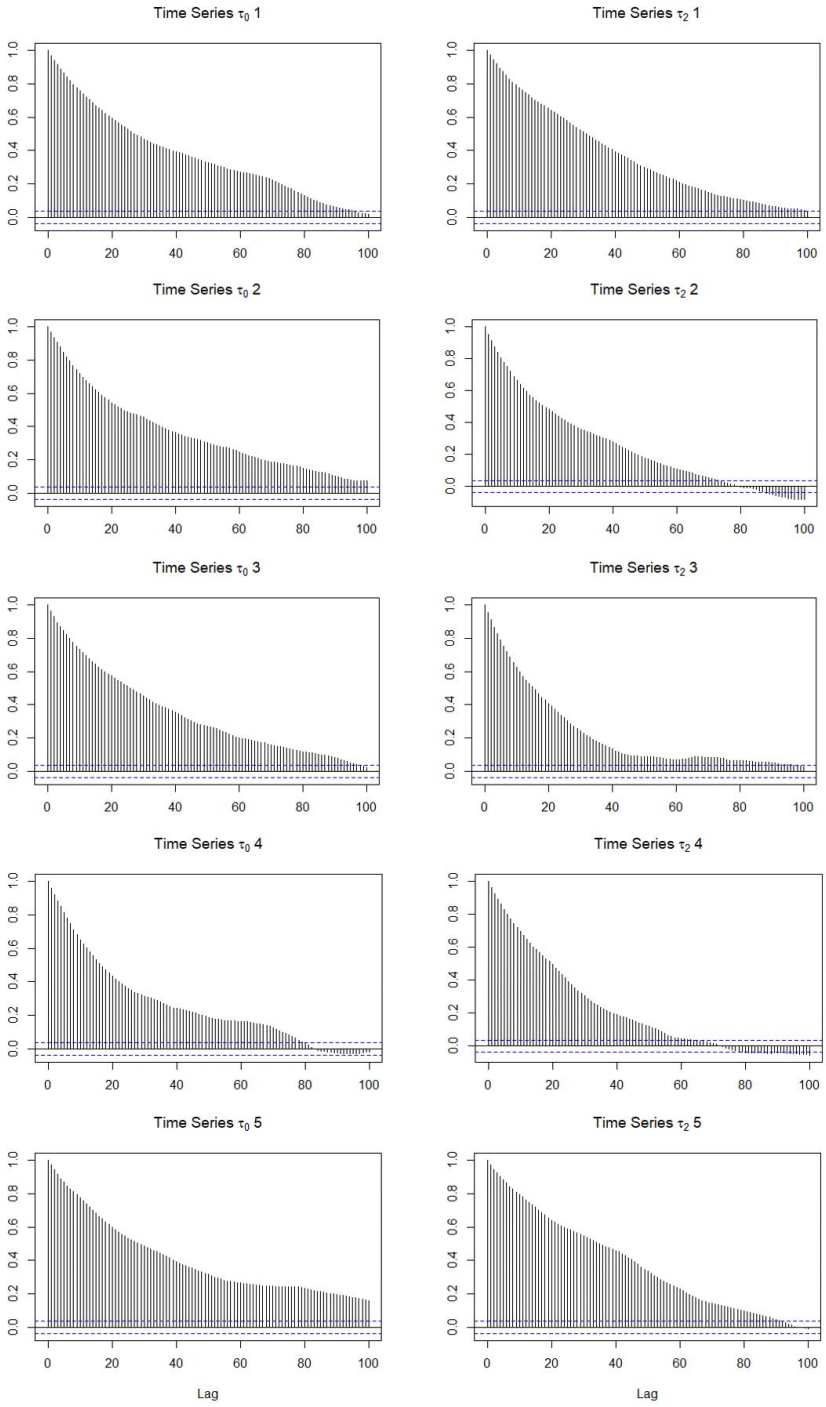

Figure A.7: Time-series plots for 5 randomly chosen τ_{0j} and τ_{2j} under the MRD model, sample size = 4,000, number of clusters = 200

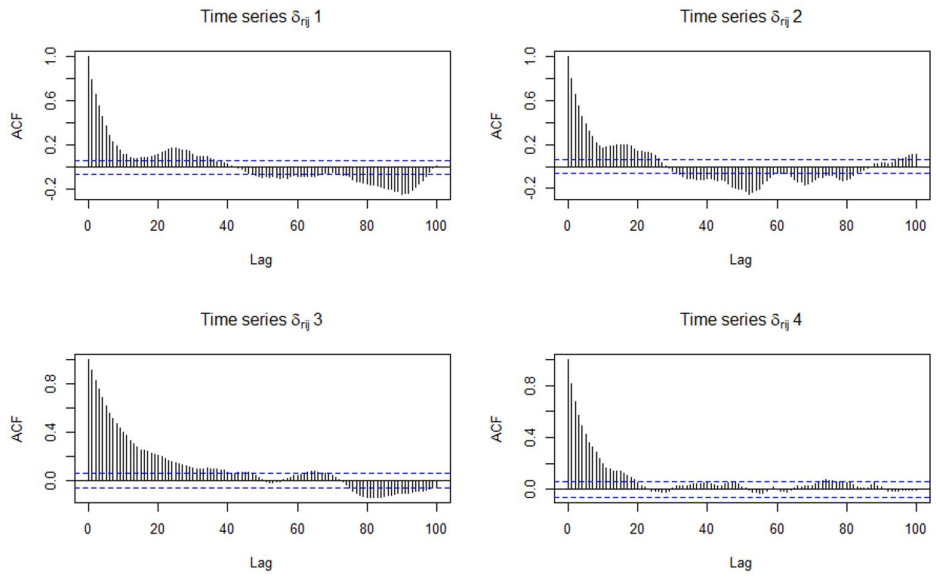

Figure A.8: Time-series plots for 4 randomly chosen δ_{rij} under MRD model with empirical data

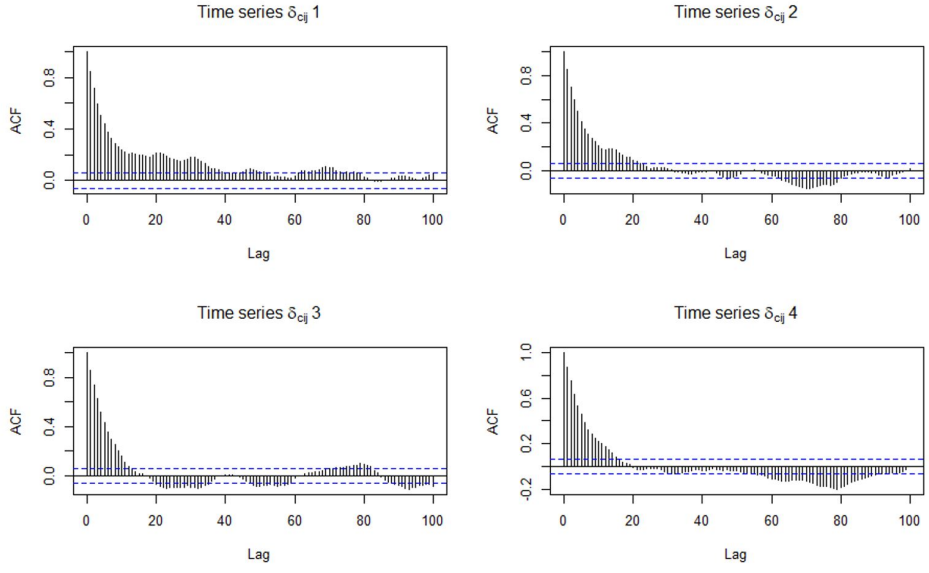

Figure A.9: Time-series plots for 4 randomly chosen δ_{cij} under MRD model with empirical data

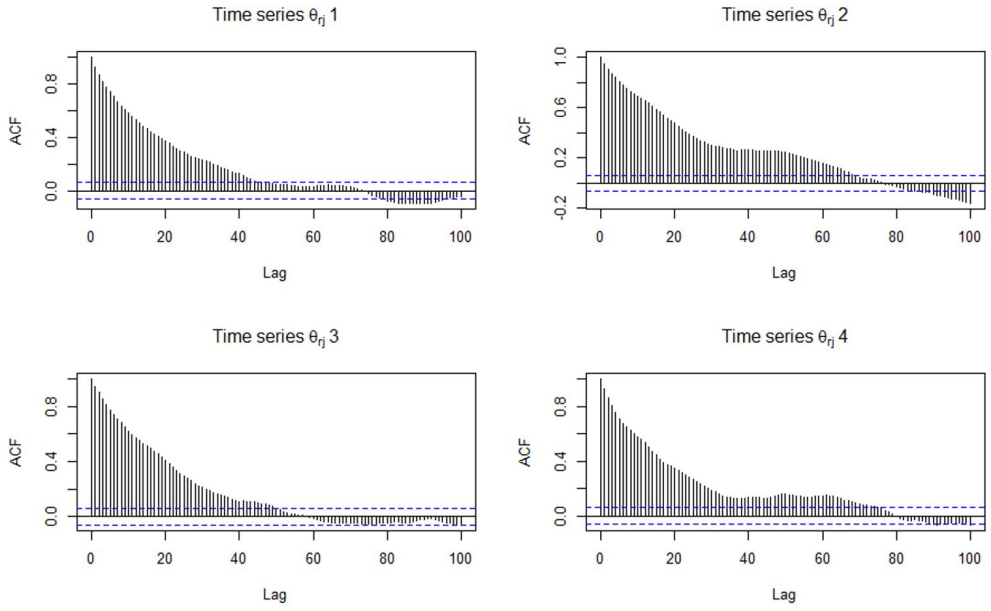

Figure A.10: Time-series plots for 4 randomly chosen θ_{rj} under the MRD model with empirical data

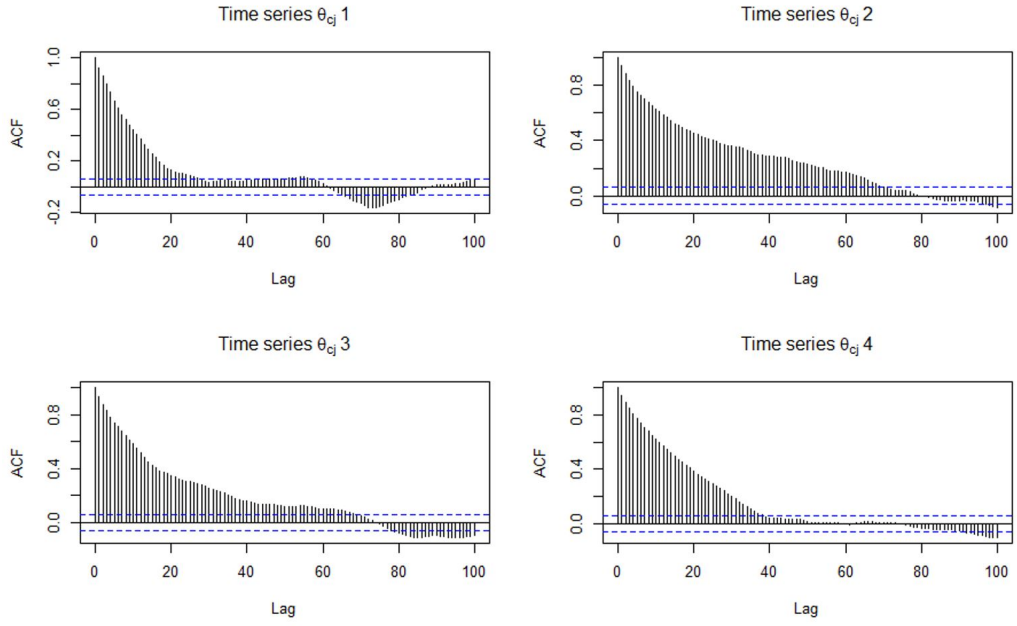

Figure A.11: Time-series plots for 4 randomly chosen θ_{cj} under the MRD model with empirical data

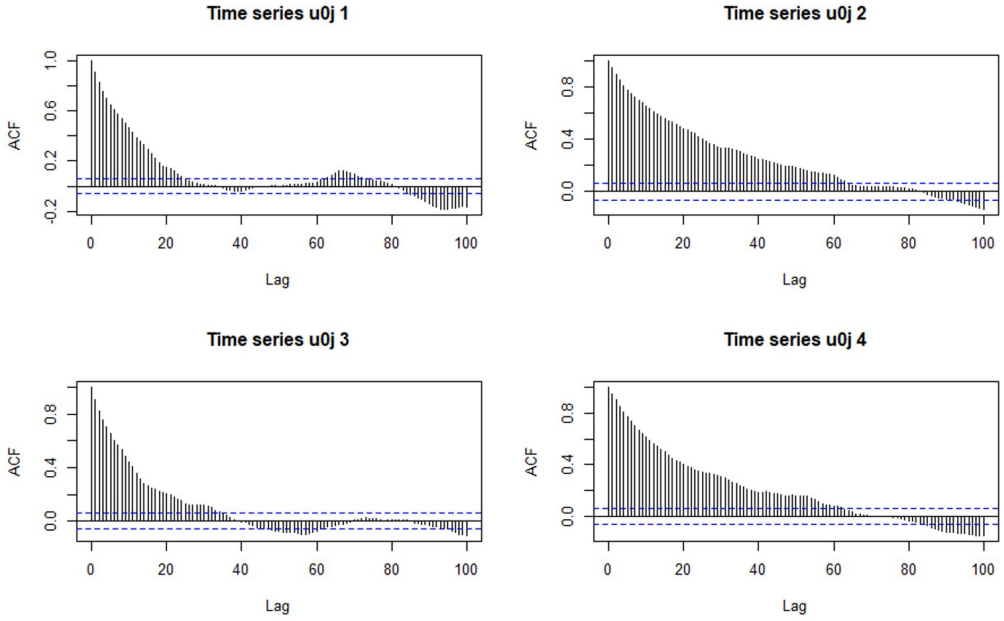

Figure A.12: Time-series plots for 4 randomly chosen u_{0j} under the MRD model using empirical data

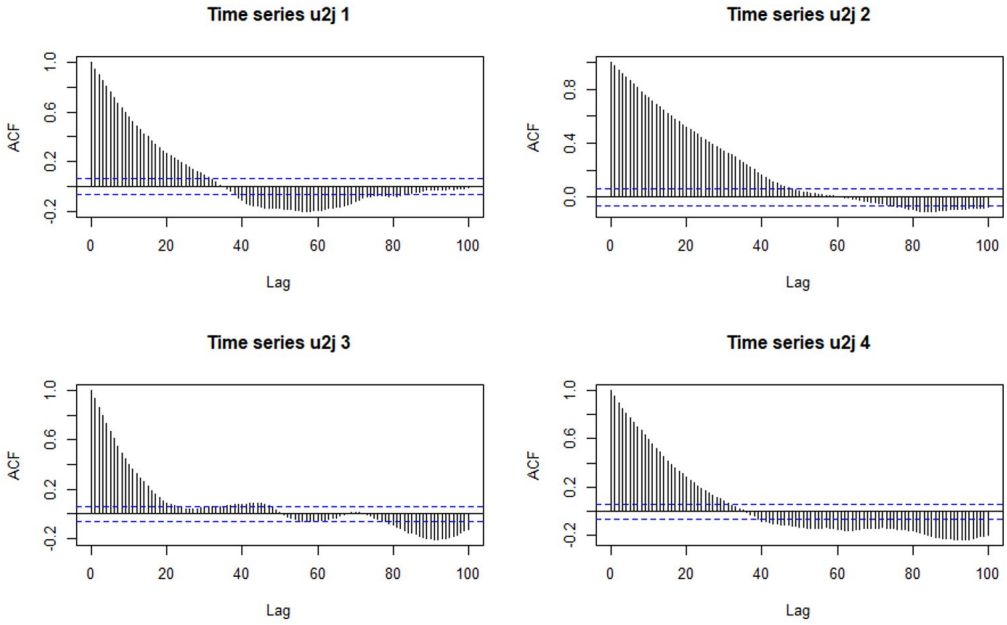

Figure A.13: Time-series plots for 4 randomly chosen u_{2j} under the MRD model using empirical data

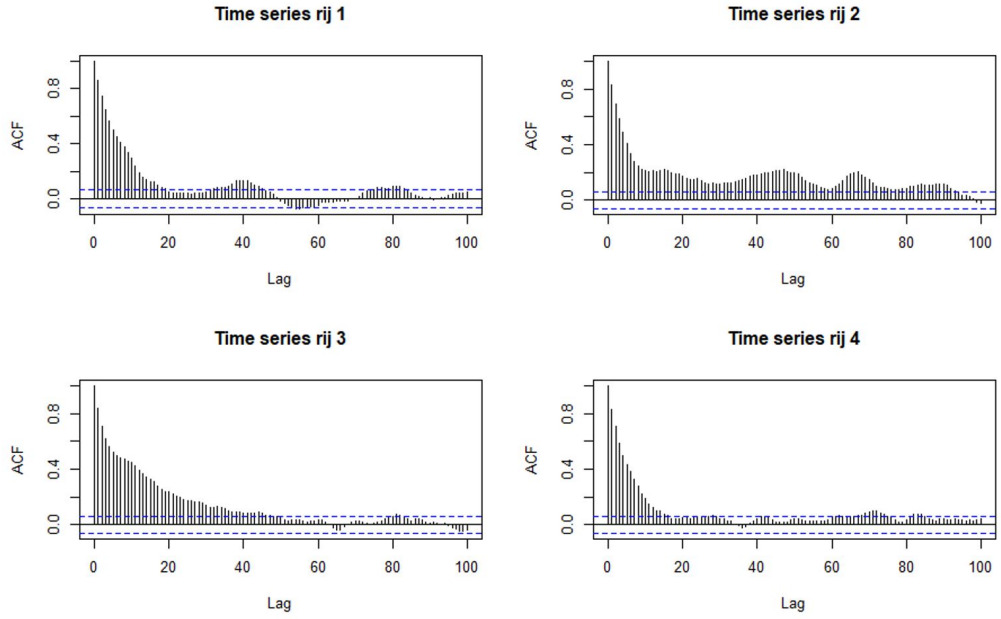

Figure A.14: Time-series plots for 4 randomly chosen ε_{ij} under the MRD model using empirical data

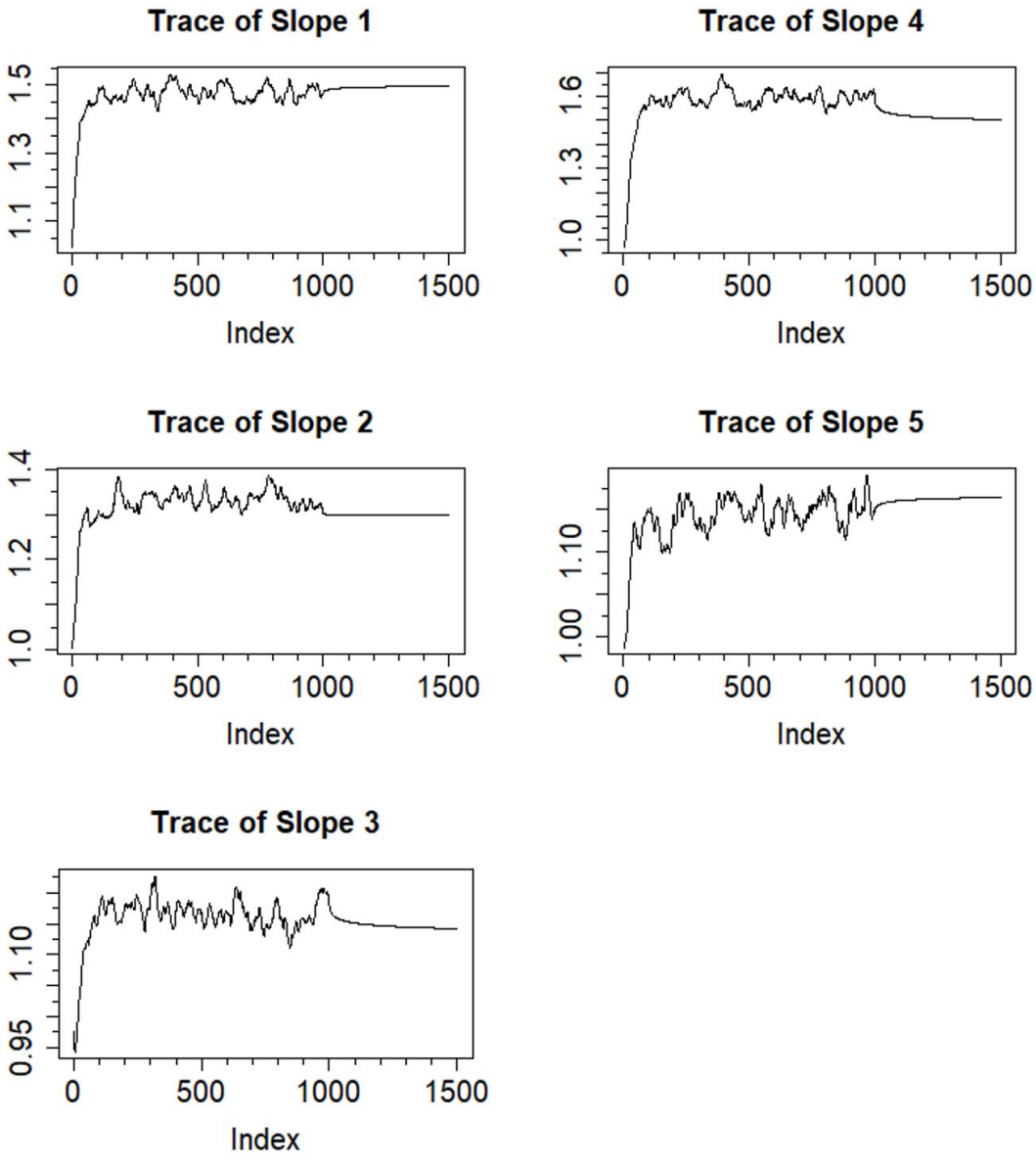

Figure A.15: Trace plots for item slope parameters, a_k, for the running variable under the HRD model, sample size = 4,000, number of clusters = 200, burn-in = 5

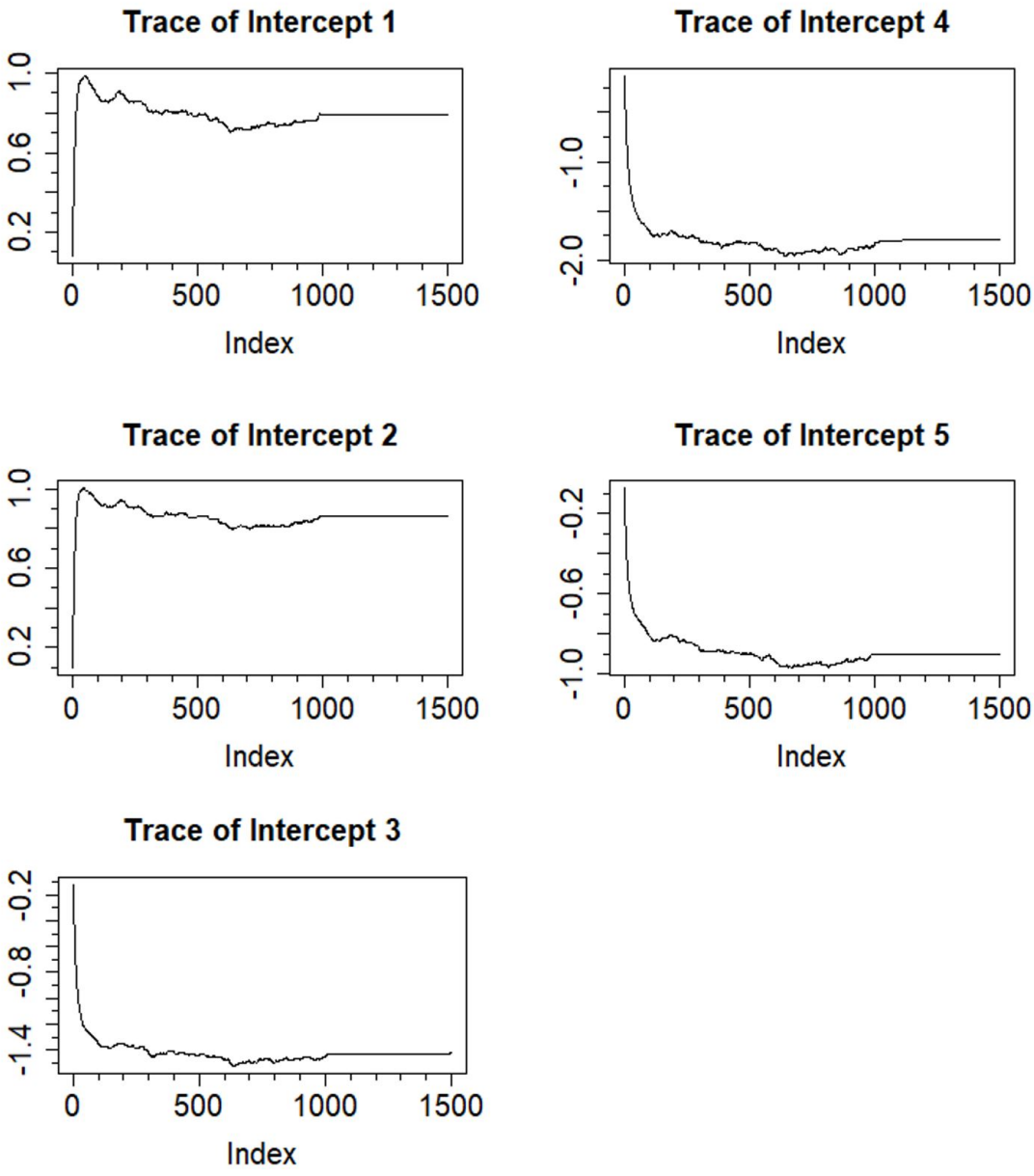

Figure A.16: Trace plots for item intercept parameters, c_k, for the running variable under the HRD model, sample size = 4,000, number of clusters = 200, burn-in = 5

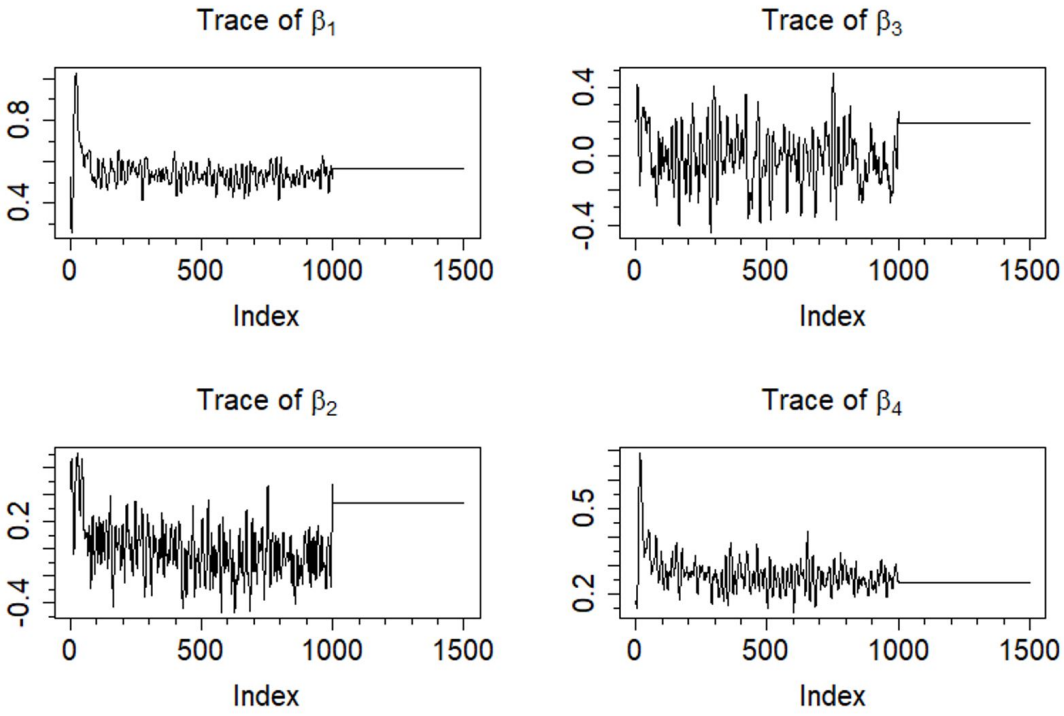

Figure A.17: Trace plots for regression slopes under the HRD model, sample size = 4,000, number of clusters = 200, burn-in = 5

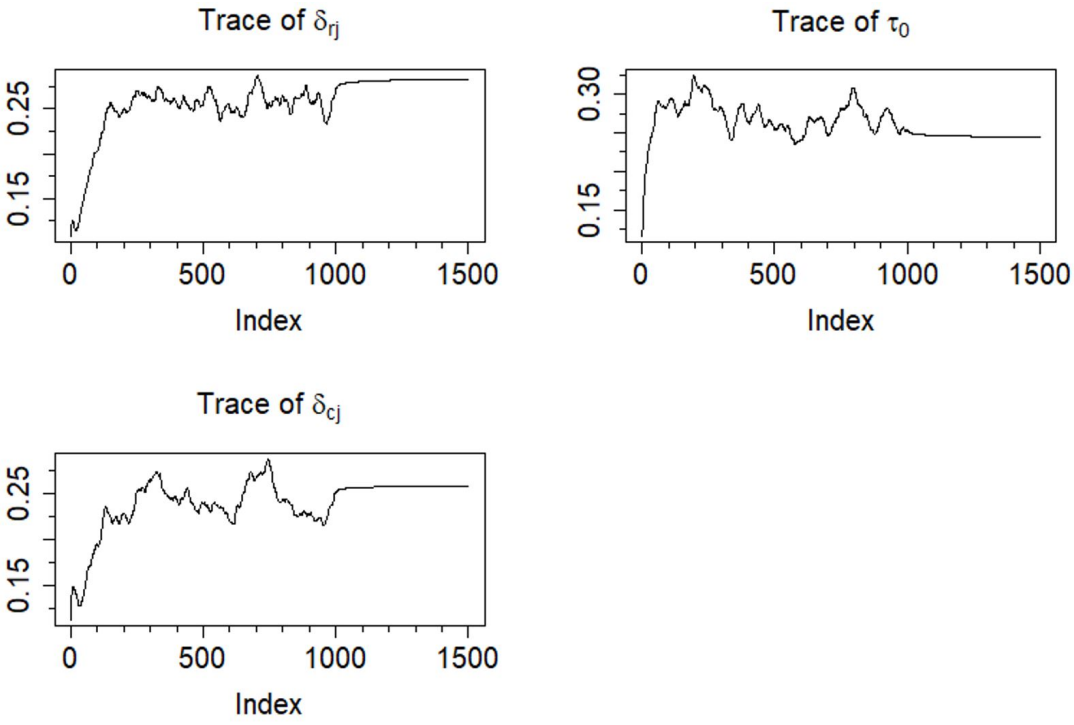

Figure A.18: Trace plots for variance parameters under the HRD model, sample size = 4,000, number of clusters = 200, burn-in = 5

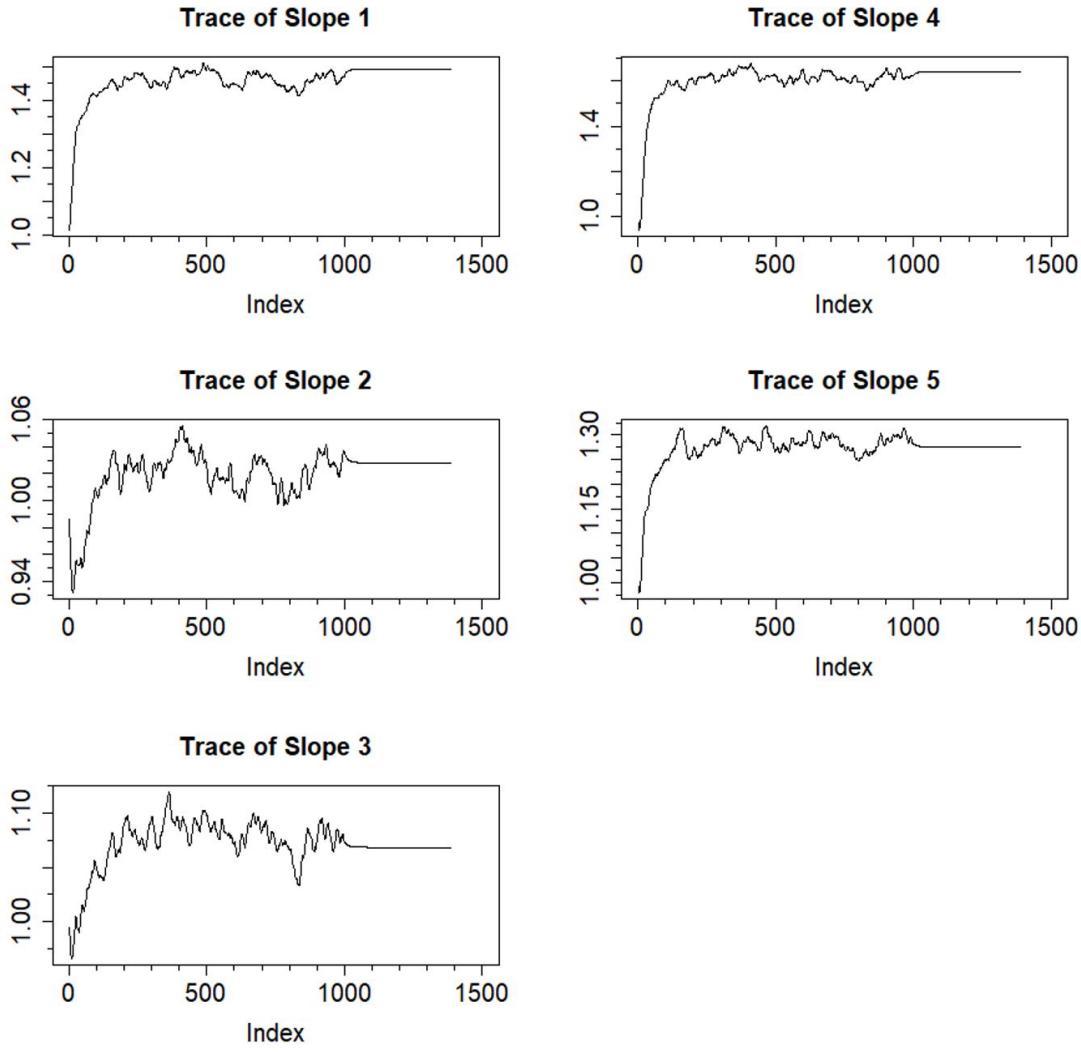

Figure A.19: Trace plots for item slope parameters, a_k, for the running variable under the MRD model, sample size = 4,000, number of clusters = 200, burn-in = 20

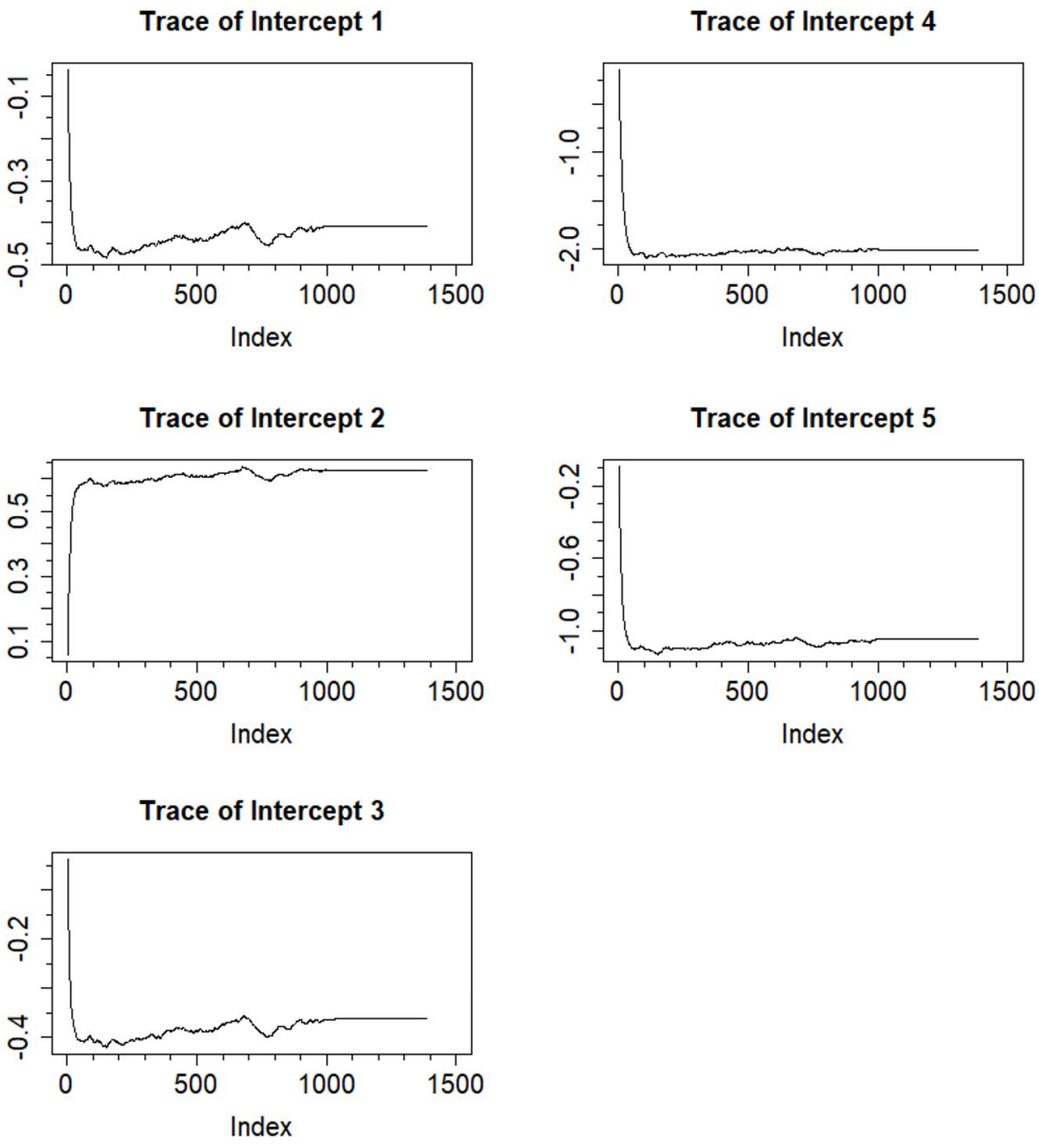

Figure A.20: Trace plots for item intercept parameters, c_k, for the running variable under the MRD model, sample size = 4,000, number of clusters = 200, burn-in = 20

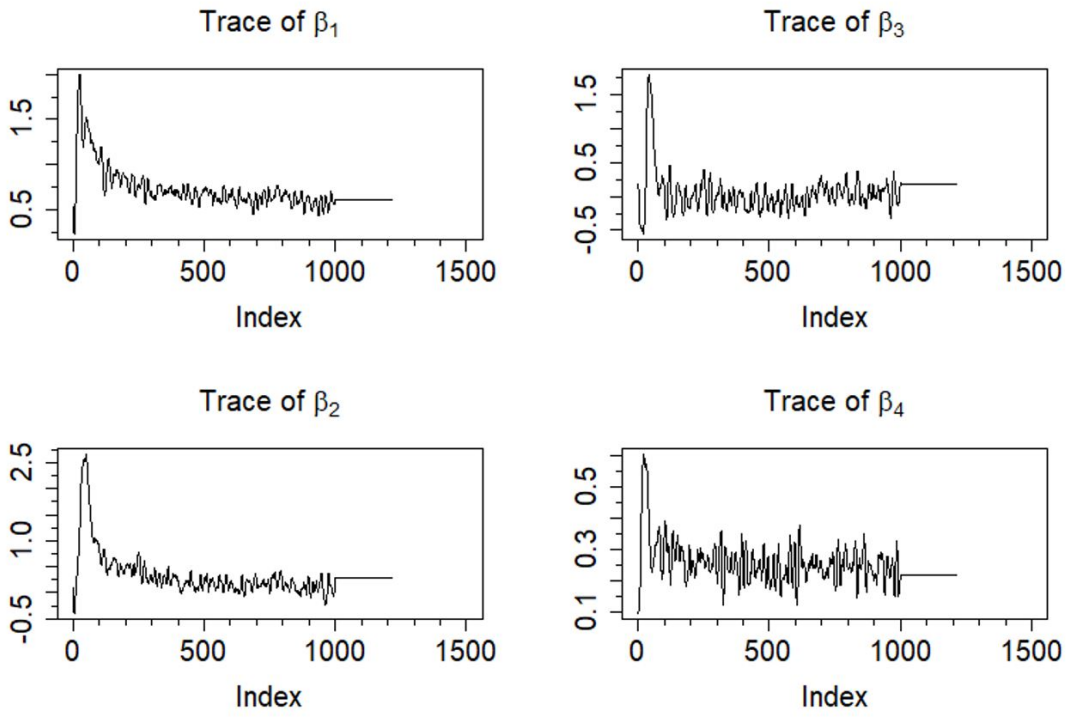

Figure A.21: Trace plots for regression slopes under the MRD model, sample size = 4,000, number of clusters = 200, burn-in = 20

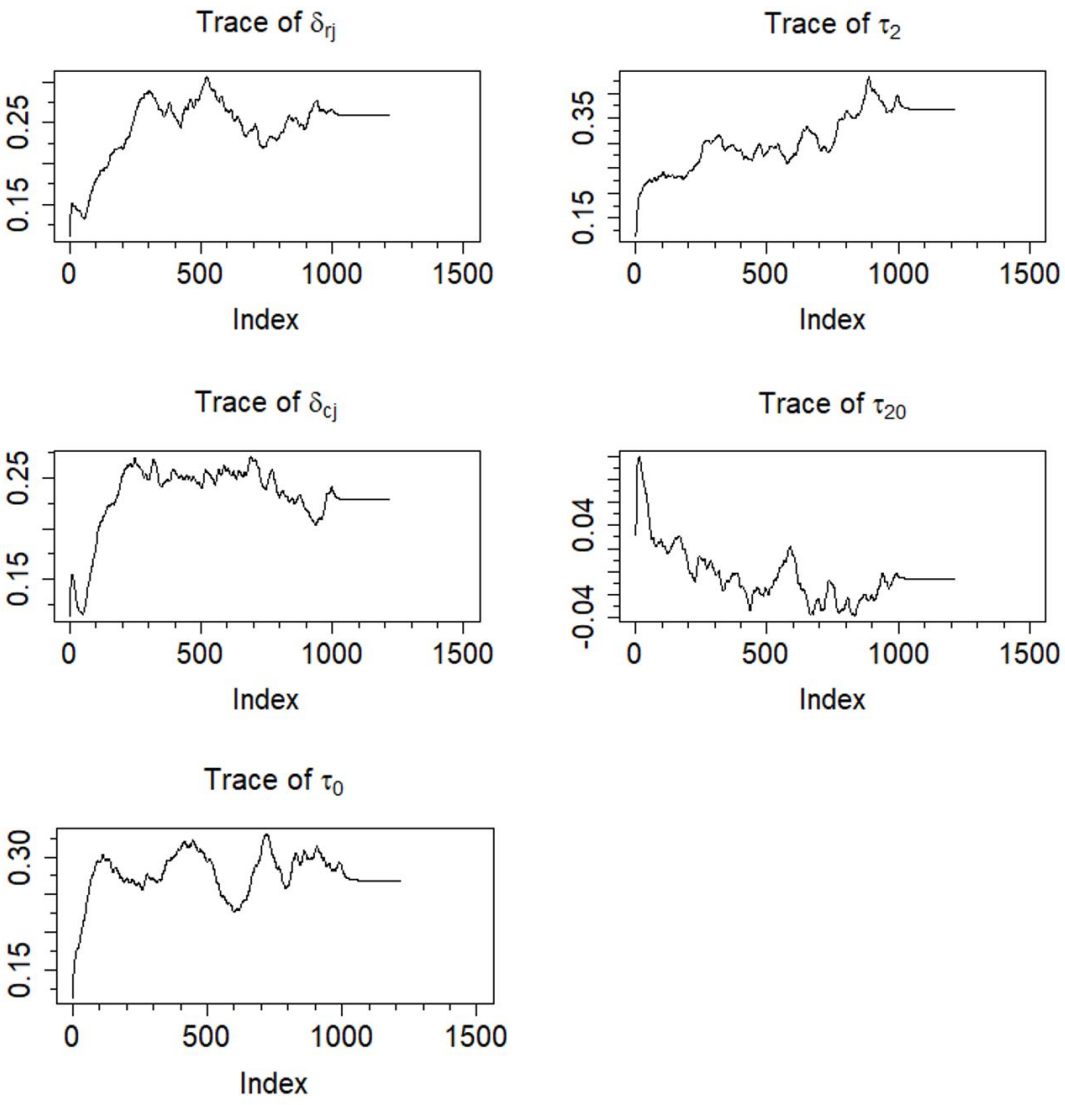

Figure A.22: Trace plots for variance parameters under the MRD model, sample size = 4,000, number of clusters = 200, burn-in = 20

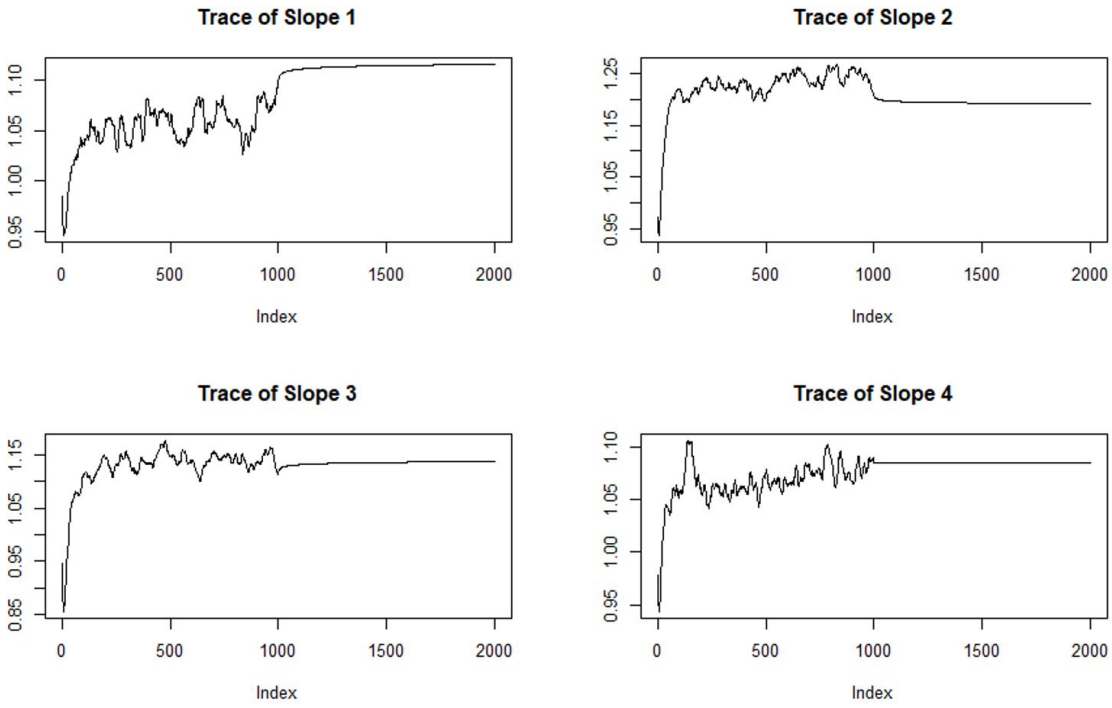

Figure A.23: Trace plots for 4 randomly chosen item slope parameters under the MRD model using empirical data

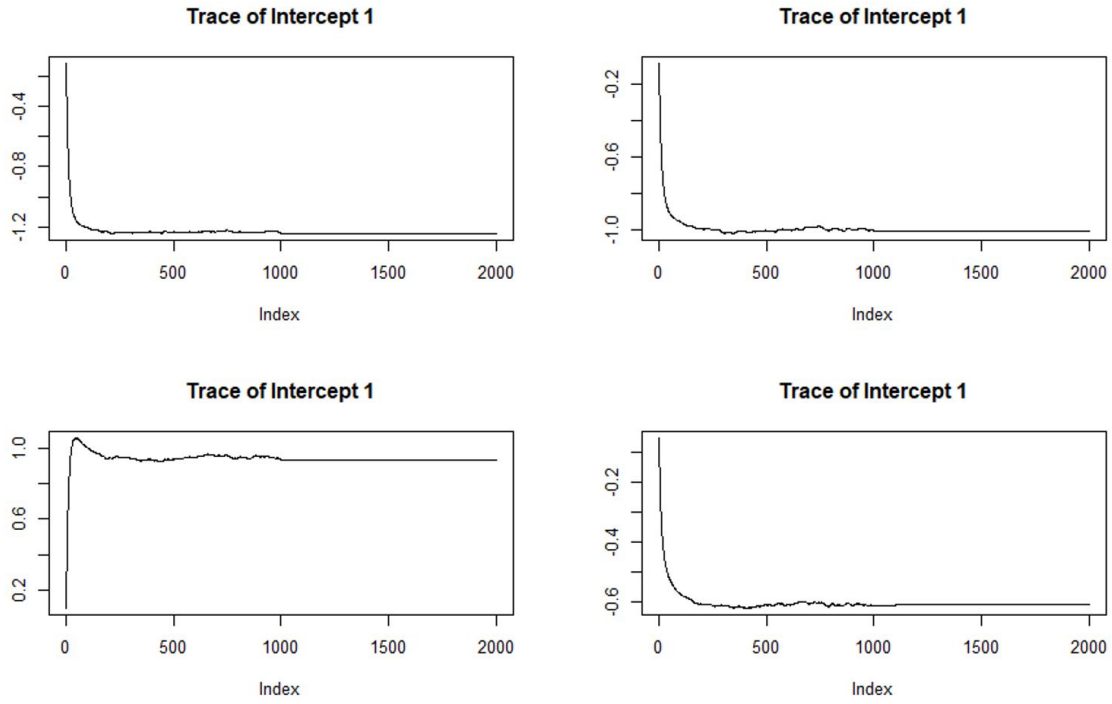

Figure A.24: Trace plots for 4 randomly chosen item intercept parameters under the MRD model using empirical data

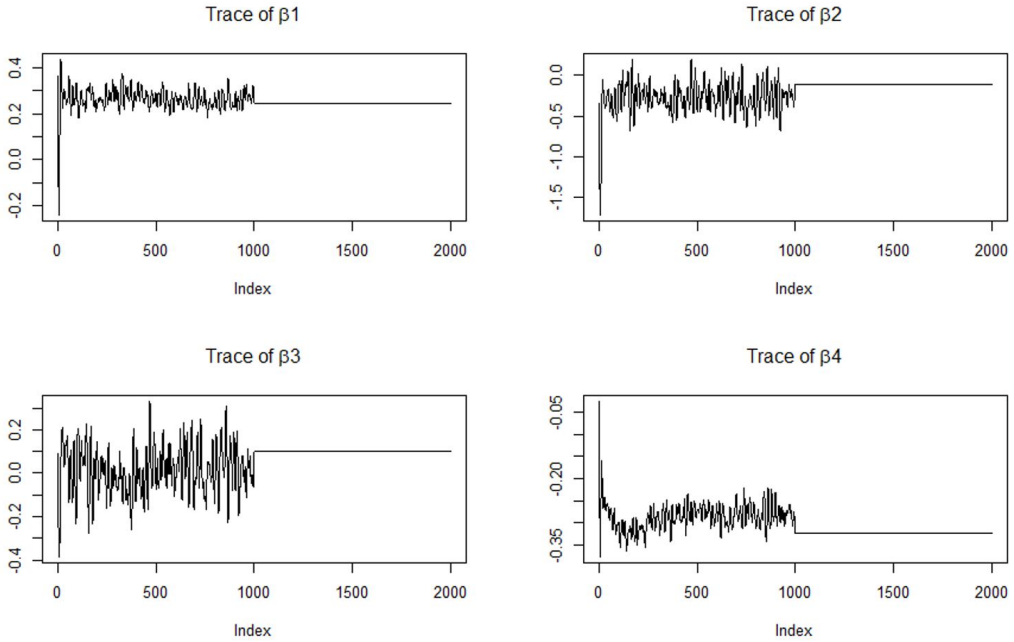

Figure A.25: Time-series plots for beta parameters under the MRD model using empirical data

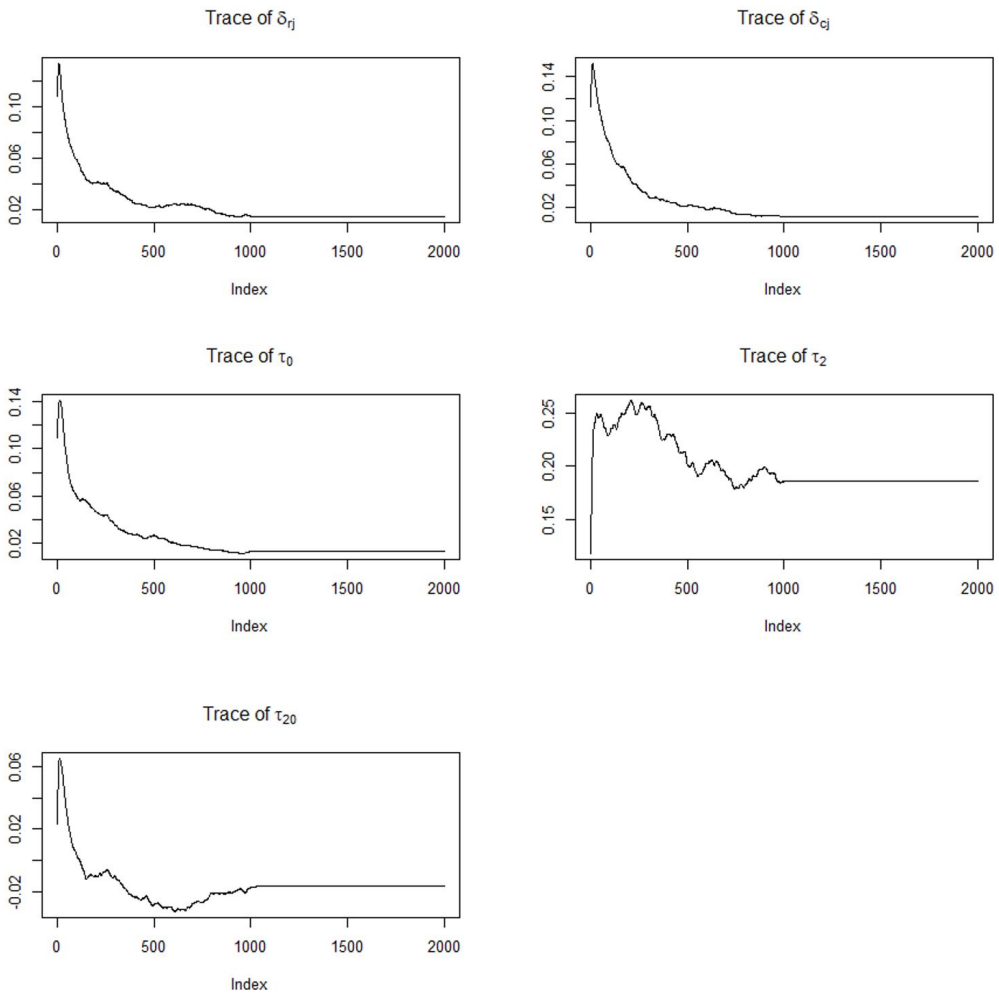

Figure A.26: Time-series plots for variance parameters under the MRD model using empirical data

Table A.7: Runtimes (in seconds) for HRD and MRD models in Simulation Study II

HRD Model									
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Average	Range	Average	Range	Average	Range	Average	Range
20	200	304	223 - 421	610	572 - 649	186	134 - 274	534	502 - 566
	500	540	502 - 596	842	732 - 924	469	432 - 470	788	741 - 856
MRD Model									
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Average	Range	Average	Range	Average	Range	Average	Range
20	200	123	103 - 154	415	386 - 491	231	216 - 249	612	523 - 736
	500	840	795 - 899	1617	1532 - 1729	429	368 - 502	524	489 - 568
40	200	1022	981 - 1103	1431	1332 - 1529	441	374 - 536	1023	967 - 1134
	500	2094	1974 - 2206	2459	2416- 2470	960	915 - 1023	2406	2345 - 2459

Note: *np* is number of individuals per cluster; Cluster is the number of clusters.

Table A.8: Gradient Norms for Simulation Study II Conditions

HRD Model								
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Average	Range	Average	Range	Average	Range	Average	Range
20	200	0.01	0 - 0.08	<0.01	0 - 0.09	0.01	0 - 0.10	0.01	0 - 0.09
	500	<0.01	0 - 0.05	0.01	0 - 0.06	0.01	0 - 0.09	0.01	0 - 0.07
MRD Model									
		Full Sample				Bandwidth = 1			
np	Clusters	10 Items		30 Items		10 Items		30 Items	
		Average	Range	Average	Range	Average	Range	Average	Range
20	200	0.01	0 - 0.03	<0.01	0 - 0.03	0.02	<0.01 - 0.12	0.01	0 - 0.08
	500	<0.01	0 - 0.02	<0.01	0 - 0.03	0.02	<0.01 - 0.14	0.01	0 - 0.07
40	200	<0.01	0 - 0.03	<0.01	0 - 0.02	0.01	0 - 0.11	0.01	0 - 0.10
	500	<0.01	0 - 0.03	<0.01	0 - 0.03	0.01	0 - 0.08	<0.01	0 - 0.09

Note: *np* is number of individuals per cluster; Cluster is the number of clusters

References

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Sage.

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*(2), 207–244.

Akaike, H. (1987). Factor analysis and AIC In *Selected papers of Hirotugu Akaike* (pp. 371–386). Springer.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*(434), 444–455.

Angrist, J. D., & Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331–1344.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*(4), 475–496.

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142.

Banks, J., & Mazzonna, F. (2012). The effect of education on old age cognitive abilities: evidence from a regression discontinuity design. *The Economic Journal*, *122*(560), 418–448.

Bartholomew, D., & Knott, M. (1999). Latent class models and factor analysis. *Oxford University Press Inc., New York*.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*(4), 373.

Bellman, R. (1957). Dynamic programming. *Science*, *153*(3731), 34–37.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Berk, R. A., Barnes, G., Ahlman, L., & Kurtz, E. (2010). When second best is good enough: a comparison between a true experiment and a regression discontinuity quasi-experiment. *Journal of Experimental Criminology*, *6*.

Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* Guilford Press.

Binder, D. A., & Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of Survey Data*, 29–48.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical Theories of Mental Test Scores*.

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

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review*, *8*(2), 225–246.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*(1), 43–82.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., & Bos, J. M. (1997). The benefits and costs of JTPA Title II-A programs: Key findings from the national job training partnership act study. *Journal of Human Resources*, 549–576.

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

Boatman, A. (2012). Evaluating institutional efforts to streamline postsecondary remediation: The causal effects of the Tennessee developmental course redesign initiative on early student academic success. An NCPR working paper. *National Center for Postsecondary Research*.

Boatman, A., & Long, B. T. (2010). Does remediation work for all students? how the effects of postsecondary remedial and developmental courses vary by level of academic preparation. An NCPR working paper. *National Center for Postsecondary Research*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika*, *35*(2), 179–197.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research*, *17*(3), 303–316.

Branson, Z., Rischard, M., Bornn, L., & Miratrix, L. W. (2019). A nonparametric

Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, *202*, 14–30.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 111–135). Beverly Hills, CA: Sage.

Burstein, L. (1980). Chapter 4: the analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, *8*(1), 158–233.

Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model* (Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill.

Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33–57.

Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307–335.

Calcagno, J. C., & Long, B. T. (2008). *The impact of postsecondary remediation using a regression discontinuity approach: Addressing endogenous sorting and noncompliance* (Tech. Rep.). National Bureau of Economic Research.

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2016). Regression discontinuity designs using covariates. *Review of Economics and Statistics* (0).

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. *Handbook of research on teaching*,171–246.

Cappelleri, J. C., Trochim, W. M., Stanley, T., & Reichardt, C. S. (1991). Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: I. The case of no interaction. *Evaluation Review*, *15*(4),395–419.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, *3*(1), 1–24.

Cawley, G. C., & Talbot, N. L. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, *22*(19), 2348–2355.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review*, *95*(4), 1237–1258.

Chiang, H. (2009). How accountability pressure on failing schools affects student achievement. *Journal of Public Economics*, *93*(9-10), 1045–1057.

Chib, S., & Jacobi, L. (2016). Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *Journal of Applied Econometrics*, *31*(6), 1026–1047.

Cliffordson, C., & Gustafsson, J.-E. (2010). Effects of schooling and age on performance in mathematics and science: a between-grade regression discontinuity design with instrumental variables applied to Swedish TIMSS 95 data. In *The 4th IEA International Research Conference, Department of Education, University of Gothenburg, from 1 July to 3 July 2010.*

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175–199.

Cook, T. D. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Economet-*

rics, *142*(2), 636–654.

Cook, T. D., Campbell, D. T., & Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston.

De Ayala, R. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Dimmery, D. (2016). rdd: Regression discontinuity estimation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rdd (R package version 0.57)

Ding, W., & Lehrer, S. F. (2007). Do peers affect student achievement in china's secondary schools? *The Review of Economics and Statistics*, *89*(2), 300–312.

Efron, B. (1960). Tibshirani,(1993), an introduction to the bootstrap. *Monographs on Statistics and Applied Probability. Chapman & Hall/CRC*, *1*.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., & Snyder Jr, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, *59*(1), 259–274.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford press.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological methods*, *12*(2), 121.

ESSA. (2015). *Every student succeeds act of 2015, pub. l. no. 114-95 114 stat. 1177 (2015-2016).*

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests.* ERIC.

Fisher, R. (1925). Theory of statistical estimation. , *22*, 700-725.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*(1), 69–78.

Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*(2), 271–288.

Frölich, M. (2007a). *Regression discontinuity design with covariates* (Tech. Rep.). IZA Discussion Paper No. 3024.

Frölich, M. (2007b). Regression discontinuity design with covariates.

Gamse, B. C., Bloom, H. S., Kemple, J. J., & Jacob, R. T. (2008). Reading first impact study: Interim report. NCEE 2008-4016. *National Center for Education Evaluation and Regional Assistance*.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision* (pp. 564–584). Elsevier.

Geneletti, S., Ricciardi, F., O'Keeffe, A. G., & Baio, G. (2019). Bayesian modelling for binary outcomes in the regression discontinuity design. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *182*(3), 983–1002.

Goldberger, A. S. (1972). *Selection bias in evaluating treatment effects: The case of interaction*. Institute for Research on Poverty, University of Wisconsin.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons.

Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 861–869.

Goodman, J. (2008). Who merits financial aid?: Massachusetts' Adams scholarship. *Journal of Public Economics*, *92*(10-11), 2121–2131.

Gu, M. G., & Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences*, *95*(13), 7270–7274.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, 193– 211.

Hahn, J., Todd, P., & Van der Klaauw, W. (1999). *Evaluating the effect of an antidiscrimination law using a regression-discontinuity design* (Tech. Rep.). National Bureau of Economic Research Working Paper 7131.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*. Chapman and Hall/CRC.

Hodara, M. (2012). *Language minority students at community college: How do developmental education and English as a second language affect their educational outcomes?* (Unpublished doctoral dissertation). Teachers College.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In *Rasch models* (pp. 53–68). Springer.

Holbein, J. B., & Ladd, H. F. (2017). Accountability pressure: Regression discontinuity estimates of how no child left behind influenced student behavior. *Economics of Education Review*, *58*, 55–67.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, *81*(396), 945–960.

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika*, *55*(1), 5–18.

Houts, C. R., & Cai, L. (2015). Flexmirt: Flexible multilevel item factor analysis and test scoring users manual version 3.0. *Vector Psychometric Group, LLC, Chapel Hill*.

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

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of economic studies*, *79*(3), 933–959.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research*, *25*(4), 467–478.

Jacob, B. A., & Lefgren, L. (2004a). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, *39*(1), 50–79.

Jacob, B. A., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity analysis. *Review of economics and statistics*, *86*(1), 226–244.

Jacob, R., Zhu, P., Somers, M.-A., & Bloom, H. (2012). A practical guide to regression discontinuity. *MDRC*.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*(2), 219.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika*, *75*(3), 393–419.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*(1), 79–93.

Kaplan, D., Kim, J.-S., & Kim, S.-Y. (2009). *Multilevel latent variable modeling: Current research and recent developments.* Sage Publications Ltd.

Karabatsos, G., & Walker, S. G. (2015). A Bayesian nonparametric causal model for regression discontinuity designs. In *Nonparametric Bayesian inference in biostatistics* (pp. 403–421). Springer.

Kelley, J., Evans, M., Lowman, J., & Lykes, V. (2017). Group-mean-centering independent variables in multi-level models is dangerous. *Quality & Quantity*, *51*(1), 261–283.

Kenny, D. A., Kashy, D. A., Bolger, N., et al. (1998). Data analysis in social psychology. *The handbook of social psychology*, *1*(4), 233–265.

Kreft, I. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Sage.

Lalive, R. (2007). Unemployment benefits, unemployment duration, and post-unemployment jobs: A regression discontinuity approach. *American Economic Review*, *97*(2), 108–112.

Lee, D. S. (2008). Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, *142*(2), 675–697.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of economic literature*, *48*(2), 281–355.

Lee, H., & Munk, T. (2008). Using regression discontinuity design for program evaluation. In *Proceedings of the 2008 joint statistical meeting* (pp. 3–7).

Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via em type algorithms. *Statistica Sinica*, 749–766.

Li, F., Mattei, A., & Mealli, F. (2015a). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, *9*(4), 1906–1931.

Li, F., Mattei, A., & Mealli, F. (2015b). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 1906–1931.

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

Lim, R. L. (1993a). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Lim, R. L. (1993b). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Linden, A., Adams, J. L., & Roberts, N. (2006). Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *Journal of Evaluation in Clinical Practice*, *12*(2), 124–131.

Littell, R. C., Henry, P., & Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*, *76*(4), 1216–1231.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*(3), 624–629.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, *49*(4), 354–371.

Liu, Y., Yang, J. S., & Maydeu-Olivares, A. (2019a). Restricted recalibration of item response theory models. *Psychometrika*, *84*(2),529–553.

Liu, Y., Yang, J. S., & Maydeu-Olivares, A. (2019b). Restricted recalibration of item response theory models. *Psychometrika*, *84*(2),529–553.

Lord, F. (1974). Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika*, *39*(2), 247–264.

Lord, F. (1975). Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters 1. *ETS Research Bulletin Series*, *1975*(2), i–52.

Lord, F. (1980). Applications of item response theory to practical testing problems. *J Lawrence Erlbaum Associates*.

Lord, F., & Novick, M. (1968). *Statistical theories of merital test scores.* Addison-Wesley Publishing Co. Reading, MA.

Lord, F., & Wingersky, M. S. (1984). Comparison of irt true-score and equipercentile observed-score" equatings". *Applied Psychological Measurement*, *8*(4), 453–461.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 226–233.

Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding irt in structural equation models: A comparison with regression based on irt scores. *Structural Equation Modeling*, *12*(2), 263–277.

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008a). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological methods*, *13*(3), 203.

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

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling*, *16*(2), 191–225.

Martorell, P., & McFarlin Jr, I. (2011). Help or hindrance? the effects of college remediation on academic and labor market outcomes. *The Review of Economics and Statistics*, *93*(2), 436–454.

Matsudaira, J. D. (2008). Mandatory summer school and student achievement. *Journal of Econometrics*, *142*(2), 829–850.

Matta, R., Ribas, R. P., Sampaio, B., & Sampaio, G. R. (2016). The effect of age at school entry on college admission and earnings: a regression-discontinuity approach. *IZA Journal of Labor Economics*, *5*(1), 9.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713.

Maydeu-Olivares, A., & Liu, Y. (2015). Item diagnostics in multivariate discrete data. *Psychological Methods*, *20*(2), 276.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, *142*(2), 698–714.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117.

McNeish, D. (2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological methods*, *19*(4), 552.

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

Melguizo, T., Bos, J. M., Ngo, F., Mills, N., & Prather, G. (2016). Using a regression discontinuity design to estimate the impact of placement decisions in developmental math. *Research in Higher Education*, *57*(2), 123–151.

Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*(435), 1254–1267.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Mislevy, R. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359–381.

Mislevy, R. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, *80*(392), 993–997.

Mislevy, R. (1992). Linking educational assessments: Concepts, issues, methods, and prospects.

Mislevy, R., Johnson, E. G., & Muraki, E. (1992). Chapter 3: Scaling procedures in NAEP. *Journal of Educational Statistics*, *17*(2), 131–154.

Mislevy, R., & Sheehan, K. (1983). Marginal estimation procedures. *The NAEP*, *1984*, 293–360.

Morell, M., & Yang, J. S. (2016). *Scoring options for assignment variable in regression discontinuity designs.* (Presented at the International Meeting of the Psychometric Society. Ashville, NC)

Morell, M., & Yang, J. S. (2017). *Using item response models to handle measurement error in regression discontinuity designs.* (Presented at the annual meeting of the National Council on Measurement in Educaiton. San Antonio, TX)

Morell, M., & Yang, J. S. (2018). *Scoring options for assignment variable in regression discontinuity designs.* (Presented at the annual meeting of the National Council on Measurement in Education. New York, NY)

Morell, M., Yang, J. S., & Liu, Y. (2019). Randomized cluster regression discontinuity design with latent variables. Presented at the 2019 Annual Learning, Educational Achievement, and Life Course Development Conference, Tubingen, Germany.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.

Muthén, B. O. (2002). Beyond sem: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Naylor, J. C., & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *31*(3), 214–225.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics* (pp. 123–150). Springer.

Nomi, T., & Allensworth, E. (2009). "Double-dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*(2), 111–148.

Ou, D. (2010). To leave or not to leave? a regression discontinuity analysis of the impact of failing the high school exit exam. *Economics of Education Review*, *29*(2), 171–186.

Paccagnella, O. (2006). Centering or not centering in multilevel models? the role of the group mean and the assessment of group effects. *Evaluation review*, *30*(1), 66–85.

Parke, W. R. (1986). Pseudo maximum likelihood estimation: The asymptotic distribution. *The Annals of Statistics*, 355–357.

Rabe-Hesketh, S., & Skrondal, A. (2004a). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Rabe-Hesketh, S., & Skrondal, A. (2004b). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Rau, T. (2011). Bayesian inference in the regression discontinuity model. *Vigesimosextas Jornadas Anuales de Economıa*.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, *18*(4), 321–349.

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

Rhoads, C. H., & Dye, C. (2016). Optimal design for two-level random assignment and regression discontinuity studies. *The Journal of Experimental Education*, *84*(3), 421–448.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*(3), 400-407.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, *38*(2), 337–341.

Rokkanen, M. A. (2015). *Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design* (Tech. Rep.). Columbia University, Department of Economics, New York, NY, Discussion Paper 1415-03.

Rosenbaum, J. E. (1995). Changing the geography of opportunity by expanding residential choice: Lessons from the gautreaux program. *Housing Policy Debate*, *6*(1), 231–269.

RStudio Team. (2018). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

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

Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, *2*(1), 1–26.

Rubin, D. B., & Thayer, D. T. (1983). More on EM for ML factor analysis. *Psychometrika*, *48*(2), 253–257.

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1).

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph No. 17*. (Richmond, VA: Psychometric Society.)

Schochet, P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*(1), 62-87.

Schwarz, G., et al. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461–464.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Seltzer, M. (2004). The use of hierarchical models in analyzing data from experiments and quasi-experiments conducted in field settings. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (p. 259-2280). Sage.

Shadish, W. R. (2011). Randomized controlled studies and alternative designs in outcome studies: Challenges and opportunities. *Research on Social Work Practice*, *21*(6), 636–643.

Shadish, W. R., Galindo, R., Wong, V. C., Steiner, P. M., & Cook, T. D. (2011). A randomized experiment comparing random and cutoff-based assignment. *Psychological methods*, *16*(2), 179.

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

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*(4), 563–575.

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

Sojourner, A. J., Frandsen, B. R., Town, R. J., Grabowski, D. C., & Chen, M. M. (2015). Impacts of unionization on quality and productivity: Regression discontinuity evidence from nursing homes. *ILR Review*, *68*(4), 771–806.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, *15*(1), 72–101.

Stapleton, L. M., McNeish, D., & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist*, *51*(3-4), 317–330.

Stigler, M., & Quast, B. (2015). rddtools: Toolbox for regression discontinuity design ('rdd') [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=rddtools (R package version 0.4.0)

Tang, Y., Cook, T. D., Kisbu-Sakarya, Y., Hock, H., & Chiang, H. (2017). The comparative regression discontinuity (crd) design: An overview and demonstration of its performance relative to basic rd and the randomized experiment. In *Regression discontinuity designs: Theory and applications* (pp. 237–279). Emerald Publishing Limited.

Team, R. C., et al. (2018). R: A language and environment for statistical computing.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In *Test scoring* (pp. 85–152). Routledge.

Thissen, D., & Steinberg, L. (2009). Item response theory. In R. Millsap & A. Maydeu-Olivares (Eds.), *The sage handbook of quantitative methods in psychology* (pp. 148 - 177). London: Sage Publications.

Thissen, D., & Wainer, H. (2001). *Test scoring*. Routledge.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309.

Trochim, W. M., & Cappelleri, J. C. (1992). Cutoff assignment strategies for enhancing randomized clinical trials. *Controlled Clinical Trials*, *13*(3), 190–212.

Tuckwiller, E. D., Pullen, P. C., & Coyne, M. D. (2010). The use of the regression discontinuity design in tiered intervention research: A pilot study exploring vocabulary instruction for at-risk kindergarteners. *Learning Disabilities Research & Practice*, *25*(3), 137–150.

Vandenbroucke, J. P., & Le Cessie, S. (2014). Commentary: regression discontinuity design: let's give it a try to evaluate medical and public health interventions. *Epidemiology*, *25*(5), 738–741.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *International Economic Review*, *43*(4), 1249–1287.

Van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, *142*(2), 731–756.

Van der Vaart, A. W., & Wellner, J. A. (1996). The delta-method. In *Weak convergence and empirical processes* (pp. 372–400). Springer.

Venkataramani, A. S., Bor, J., & Jena, A. B. (2016). Regression discontinuity designs in healthcare research. *BMJ*, *352*, i1216.

Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *The Annals of Mathematical Statistics*, *18*(1), 88–96.

Wang, C., Weiss, D. J., & Su, S. (2019). Modeling response time and responses in multidimensional health measurement. *Frontiers in Psychology*, *10*, 51.

Wong, V. C., Steiner, P. M., & Cook, T. D. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, *38*(2), 107–141.

Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis–Hastings Robbins–Monro algorithm. *Journal of Educational and Behavioral Statistics*, *39*(6), 550–582.

Yang, J. S., & Seltzer, M. (2015). Handling measurement error in predictors using a multilevel latent variable plausible values approach. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 295–333.