Quasi-Experimental Evaluations with Multi-Item Outcomes: Potential Problems and Solutions

A Dissertation

Presented to

The Faculty of the School of Education and Human Development

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Eli Talbert

May 2023

Table of Contents

Dedication

I dedicate this dissertation to all the people that have supported me along the way including my parents, Nancy and Joe Talbert, my advisors, Jim Soland and Vivian Wong, and the people that I spent most of my non-work time in graduate school with Jessica Tennis and Diana Burden. I am also grateful for all of my fellow graduate students both past and *present* who provided an intellectual community at UVA. Finally, I would like to thank the institutions that provided monetary support of various types that allowed me to complete my degree in financial security.

Overall Dissertation Abstract

This dissertation seeks to investigate measurement problems related to using multi-item measures in two classes of quasi-experimental designs. It features three chapters with each chapter functioning as an independent paper. The first paper focuses on the Difference-in-Difference design and uses simulation to outline the possible effects of sum scoring and measurement invariance when using survey scale like measures. The second paper shows how this works empirically by rescoring an already completed paper with different IRT methods and showcasing the resultant differences. Finally, the third paper investigates the viability of using a measurement model to integrate information from multiple raters in RD designs. Together the three papers help improve the rigor of quasi-experimental evaluations by integrating advances in measurement with those in causal inference.

Keywords: quasi-experimental; program evaluation; measurement; causal inference; methods

Dissertation Overview

Often in education the outcomes policymakers and researchers are interested in are intangible, also known as latent, variables. These socio-emotional and cognitive outcomes are typically measured using tools that use multiple items to estimate the construct of interest. For example, survey scales, a set of related survey questions, are commonly used to measure psychological constructs like parental perceptions of school engagement (Schueler et al., 2017). Other multi-item tools include observer protocols and tests. When used in any type of analysis, including evaluations, the multiple items from these tools must be combined in some way.

Prior work by Soland, Kuhfeld, and Edwards (2022) has delineated the many measurement decisions surrounding the combination of these items that can have impacts on analysis. Unfortunately, many researchers are unaware of the importance of these decisions and default to the simplest approach, sum scoring (D. Bauer & Curran, 2015). Flake et al. (2017) found that only 21% (37 out of 177) of the studies they reviewed used a latent variable model rather than a sum score and a mere 2% of author-developed scales reported any evidence of internal structure (3 out of 124). Given the strong assumptions underlying sum scoring (McNeish & Wolf, 2020), it is not surprising that using sum scores can bias parameters in complex procedures like growth curve modeling (Kuhfeld & Soland, 2020).

Importantly, for education policy, sum scoring can also bias causal estimates. Work with experiments has shown that sum scoring can result in treatment estimates that are up to 25% understated (Soland, 2022) while in regression discontinuity designs treatment effect estimates can be understated by up to 40% (Soland, Johnson, et al., 2022) . This bias could potentially mislead policymakers into thinking a policy or intervention is ineffective. There are also other

measurement problems like differences in measurement across groups or timepoints, known as measurement noninvariance, that have been flagged as potential problems (Soland, 2021).

The already demonstrated impact of measurement on causal inference with latent variables creates a need for further examination of measurement in quasi-experimental designs. There is potential for measurement problems to manifest itself differently in form or magnitude given each quasi-experimental design's unique logic and the different types of multi-item outcomes that can be used. In particular, DiD designs might be more susceptible to measurement problems because of their reliance on the parallel trends assumption (Cunningham, 2020) and observational protocols have rater effects (Jones & Bergin, 2019; Styck et al., 2021) that introduce additional complications.

This dissertation focuses on these two issues through three studies: two related to measurement with DiD designs and one regarding the use of multi-rater observational protocols in an RD design. Both types of quasi-experimental designs are important for education policy evaluation. The DiD design is important because it is one of the most highly flexible and widely applicable tools for causal evaluation, while the RD design is important because it is a design that can exploit the frequent use of tests and ratings to place students and direct resources in education. The dissertation, as a whole, expands research on the measurement decisions that accompany using survey scales or other multi-item instruments to measure an outcome of interest in quasi-experimental designs. The goal is to outline and illustrate measurement challenges and provide guidance on the feasibility and tradeoffs of possible solutions.

Measurement in the DiD Design

The first two chapters of my dissertation cover measurement in the DiD design. The DiD design works by accounting for the differences between treated and untreated group(s) before treatment

when comparing the outcomes of those groups. Its core logic is that any differences between the treatment and non-treatment groups will remain consistent over time and adjusted for, allowing for an unbiased causal estimate (Caniglia & Murray, 2020). This logic is formalized as the parallel trends assumption (Wing et al., 2018) and makes the DiD design susceptible to measurement problems. This susceptibility springs from the inability of the classic estimating equation of a DiD model to account for measurement differences, even though the DiD design is built to account for potential group or time differences.

In the first chapter I explore these issues through Monte-Carlo simulation in a sole authored study entitled "Difference in Differences and the Impact of Multi-Item Instrument Scoring Decisions." This paper examines the effects of using a DiD with short to medium length multi-item scales for three measurement issues. These were model misspecification from using sum scores, measurement noninvariance between control and treatment groups, and measurement noninvariance caused by response shifts. Before diving into the simulations, the paper walks through the logic of how problems with scoring can arise for a substantive researcher audience. This allows readers to understand the results of the simulation better and conceptually grasp the problems of using the most popular scoring approach. The paper then examines the three measurement problems in three separate simulation studies. In each study item loadings, sample size, length of survey scale, and treatment effect were varied to create plausible conditions with a basic DiD design to illustrate how different scoring approaches might affect causal estimates.

The main finding regarding sum score bias was that sum score bias was greatest for short scales with low item loadings. When this was the case the treatment effect was biased down by approximately 40% of the true treatment effect. The main finding regarding measurement noninvariance between control and treatment groups was that CT invariance could counteract

sum score bias in some cases but made the recovery of evidence supporting the parallel trends assumption more difficult. Finally, the main finding regarding response shifts is that they can result in bias of up to 50%.

Besides these findings, the major contribution of this paper is introducing scoring as a factor that should be considered by users of the DiD designs who might have previously neglected measurement issues. Using a Monte-Carlo simulation, this paper demonstrates, in the most basic DiD design, what could happen under plausible measurement invariance, sample size, loading, and effect size conditions. This draws attention to scoring as an issue in the DiD design and builds a foundation for future methodological investigation into scoring and the DiD design.

The second chapter builds upon the simulation results of the first. I present a co-authored paper with James Soland entitled "Can Scoring Decisions Affect Results from Quasi-experimental Studies?" This paper aims to examine the practical impact of different scoring methods by rescoring a previously published study (Whitney & Candelaria, 2017) using Item Response Theory methods. The original study covers many outcomes and contains both null and significant results making the study's rescoring an appropriate illustration of how scoring methods are not merely a technical concern and how they can change substantive findings from a study. The second chapter complements the first chapter by focusing on the practical implications that different scoring approaches can have on a study's substantive conclusions.

The major contribution of the second chapter is connecting prior technical work to a real evaluation of an already implemented policy. The study found that using a scoring model that better reflected the data and study design recovered different subgroup treatment effects and more statistically significant treatment effects for the entire student population. This helps build the case that scoring is something that is both important and feasible for substantive researchers

to consider. The study recaps the decisions that are made when scoring a measure, how they have been shown to affect analysis in other studies, and the conceptual reasons they would affect measurement. Combined with the empirical results, the study makes a strong case that researchers should pick scoring methods that align with the structure of their study and data. This chapter parallels other work that shows how scoring can affect studies that use other quantitative methods (Soland, Kuhfeld, et al., 2022).

Multi-rater Observer Protocols in an RD Design

Instead of focusing solely on a single type of outcome and quasi-experimental design, the last chapter of my dissertation focuses on the conjunction of the RD design with observer protocols, specifically RD designs that use outcomes observed by multiple raters. RD designs exploit the variability around decision rules to estimate causal effects. For example, the state of North Carolina assigns quality ratings for early childcare education on a 1-5 star scale based on a continuous measure of quality (Bassok et al., 2019). Those close to either side of the cutoff of three vs. four stars are similar allowing for the estimation of a causal estimate by using those just below the cutoff as a control group and those just above as the treatment group.

However, observational protocols introduce their own problems that go beyond the measurement issues addressed in the first two papers that are more aligned with survey scores. Specifically, there are many sources of variation, such as the context of the rating or the rater's perspective, that can introduce additional variability into the produced score (Ho & Kane, 2013) . One strategy to mitigate this problem is to use multiple raters, but there is evidence that in order to be more useful than using a single rater a researcher must use a psychometrically defensible measurement model (van Dulmen & Egeland, 2011).

In the third chapter I present a simulation study named “Combining Multi-Rater Observer Protocols with Regression Discontinuity Design for Unbiased Causal Effect Estimation” that was co-authored with James Soland. This last chapter introduces a new model that combines the tri-factor model proposed by Bauer et al. (2013) called the RD multi-rater model. The study evaluates the feasibility of using the model and the potential gains in power and bias that using a properly specified SEM model can have in comparison to alternate approaches of dealing with multiple raters or using a single rater with observational protocols. The paper includes both fuzzy and sharp RD designs. Like the first chapter this study explores the impact of scoring method when using multi-item outcomes in a quasi-experimental design but focuses on a different research design and type of outcome.

The major contribution of this chapter is to expand on previous work that outlined how using SEM to estimate treatment effects on a latent variable outcome in a RD design can improve power and reduce bias (Soland, Johnson, et al., 2022), and to propose a new model that researchers can use to improve the accuracy of causal estimates in education research. In the chapter we identified the difficulties associated with using observer protocol data from multiple raters in education research, specifically when utilizing regression discontinuity (RD) designs and explained why our proposed model, the RD Multi-rater model might be a solution. Our findings indicated that the RD Multi-rater model necessitates large sample sizes to achieve sufficient power, but it is still preferable to other approaches when sub-optimally powered, as inappropriate measurement models can result in significant bias and increased Type II error rates. Moreover, our study indicated that averaging multiple ratings may lead to a false sense of accuracy, and substandard measurement may negate the benefits of increasing sample sizes in RD designs.

Together these three chapters aim to provide a base for substantive researchers to understand the impact that scoring methods can have on the evaluations that use the DiD and RD designs and provide methods to produce more accurate treatment effect estimates. This will improve the rigor of evaluations and has the potential to integrate advances in measurement with those in causal inference.

Overview References

Bassok, D., Dee, T. S., & Latham, S. (2019). The effects of accountability incentives in early childhood education. *Journal of Policy Analysis and Management*, *38*(4), 838–866.

Bauer, D., & Curran, P. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 3–38.

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475.

Caniglia, E. C., & Murray, E. J. (2020). Difference-in-Difference in the Time of Cholera: A Gentle Introduction for Epidemiologists. *Current Epidemiology Reports*, *7*(4), 203–211. https://doi.org/10.1007/s40471-020-00245-2

Cunningham, S. (2020). Causal Inference. *The Mixtape*, *1*.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, *24*(2), 91–118.

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*. https://doi.org/10.1037/met0000367

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.

Schueler, B. E., McIntyre, J. C., & Gehlbach, H. (2017). Measuring Parent Perceptions of Family-School Engagement: The Development of New Survey Tools. *School Community Journal*, *27*(2), 275–301.

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, 01466216211013102.

Soland, J. (2022). Evidence That Selecting an Appropriate Item Response Theory–Based Approach to Scoring Surveys Can Help Avoid Biased Treatment Effect Estimates. *Educational and Psychological Measurement*, 00131644211007551.

Soland, J., Johnson, A., & Talbert, E. (2022). Regression Discontinuity Designs in a Latent Variable Framework. *Psychological Methods*.

Soland, J., Kuhfeld, M., & Edwards, K. (2022). How survey scoring decisions can influence your study's results: A trip through the IRT looking glass. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000506

Styck, K. M., Anthony, C. J., Sandilos, L. E., & DiPerna, J. C. (2021). Examining rater effects on the classroom assessment scoring System. *Child Development*, *92*(3), 976–993.

van Dulmen, M. H., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedures. *International Journal of Behavioral Development*, *35*(1), 84–92.

Whitney, C. R., & Candelaria, C. A. (2017). The effects of No Child Left Behind on children's socioemotional outcomes. *AERA Open*, *3*(3), 2332858417726324.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, *39*.

Chapter 1: Difference in Differences and the Impact of Multi-Item Instrument Scoring Decisions

Chapter 1 Abstract

The Difference in Difference (DiD) design is a popular quasi-experimental technique that is used to evaluate many types of policies. Many of the outcomes evaluated using DiD are measured using multiple items, such as surveys or observer checklists. The most common way of scoring these instruments is to take the average or the sum of the items, a practice known as sum scoring. However, this scoring approach introduces additional assumptions about the underlying relationship between the items and the target outcome that may not hold. Using Monte Carlo simulation this paper demonstrates how the decision to use sum scoring can bias the causal estimate of a DiD design and affect the evidence supporting the parallel trends assumption. This paper shows how an alternate approach that utilizes structural equation modeling, an analytic tool popular in psychology, can resolve these problems and spotlights conditions where these problems may be particularly impactful.

Keywords: structural equation modeling; difference-in-difference designs; program evaluation; causal inference.

Difference in Differences and the Impact of Multi-Item Instrument Scoring Decisions

Researchers often use quasi-experimental designs to evaluate the impact of policies and interventions in situations where an experiment is impractical or impossible. In particular, Difference-in Difference (DiD) designs are a popular quasi-experimental approach because they use data that is often already collected and can be applied to evaluate policy decisions in many situations. Recent within-study comparison results have shown that DiD designs are often able to replicate effect estimates in field settings from an experimental benchmark with the same target population (Bifulco, 2012; Hallberg et al., 2020; Somers et al., 2013; St. Clair et al., 2014), and methodological work has highlighted ways to improve estimates in more complex conditions such as multiple treatment periods (Callaway & Sant'Anna, 2020) and variations in treatment timing (Goodman-Bacon, 2021). However, less attention has been given to fully exploring the potential impact of how outcomes are measured on the analysis of a DiD design.

Potential problems arising from the construction of measurement instruments are not new and have been described in the context of threats to validity. For example, Shadish, Cook, and Campbell discuss instrumentation, where the nature of a measure changes over time in a way that can be confused for a treatment effect, and give an example of how a more expansive definition of crime by the Chicago police led to an illusory increase in crime (2002). Such descriptions, however, do not provide a formal understanding of the problem that can inform researchers of: the extent of the potential threat to internal validity, diagnostic procedures, or possible solutions. This paper explores how the construction of a specific type of measurement instrument is a threat to validity. Specifically, it addresses how scoring methods of survey scales, observer checklists, and other multi-item instruments can quantitatively affect a DiD analysis.

The examination of survey-based and other multi-item instrument outcomes is of growing importance as many of the constructs education researchers are interested in cannot be measured with a single item. For example, while wages are often reported to various government agencies as a single number, education data often looks at cognitive and emotional outcomes like engagement that require survey scales to measure (Parsons & Taylor, 2011). While currently less commonly used as outcome measures in evaluations, these constructs are often linked to the internal logic of policies. For example, an early critique of school accountability policies was the potential to induce stress in students and negatively impact their achievement (Sheldon & Biddle, 1998). Whitney & Candelaria (2017) were able to test this theory using survey outcomes and a DiD design. As education moves towards understanding how and why policy changes and interventions impact students, multi-item measures will likely become increasingly popular as an outcome in quasi-experimental evaluations.

The paper builds on prior work showing that measurement problems that arise from the combination of multiple item responses can bias effect estimates in both randomized control trials (Soland, 2022) and regression-discontinuity designs (Soland, Johnson, & Talbert, 2022). This bias arises in two primary ways. The first is if the items are combined in a way that does not reflect the true relationship between the items and the outcome they are meant to measure (McNeish & Wolf, 2020). Any misspecification of the relationship between the items and the construct it is meant to measure can produce a biased score because the items are improperly weighted (Rhemtulla et al., 2019). These biased scores can, in turn, bias effect estimates. For example, mean scores are often used to score surveys, and are produced by simply averaging the item responses. Such an approach assumes that all items should be given equal weight when measuring the latent construct. This assumption, as well as several others, is strong, and often not

met. For example, a scale meant to measure science behavioral engagement (Wang et al., 2016) features both the items: “I put effort into learning science” and “I talk about science outside of class”. Such items may represent different levels science behavioral engagement, which would make treating them as equivalent inappropriate. This paper will show that incorrect scoring of outcomes can bias estimates in DiD approaches and suggest methods for addressing this issue.

The second problem is that, even if the correct approach to combining multiple items is adopted initially, the effects of the treatment or other factors can result in changed internal standards, values, or conceptualization of the construct, also known as a response shift (Oort, 2005; Sprangers & Schwartz, 1999). In turn, a response shift might mean that the scoring approach used initially is no longer appropriate, even if an approach more sophisticated than using mean scores is employed. Instead of detecting the true effect, an evaluation might detect changes in how an instrument functions because the respondent is thinking about the questions in a different way. For example, Fokemma et. al (2013) found that therapy tends to result in patients overreporting depression because they became more familiar with depression symptoms and that this biased the estimated treatment effect downward.

This paper extends prior work (Soland, 2021, 2022; Soland et al., 2022) by describing how these problems might manifest in DiD designs, and how biases are not “differenced out” over time. The paper examines how biases arising from incorrect scoring decisions of multi-item outcomes may be particularly large for DiD approaches because of the complications of using non-equivalent control and treatment groups. The different compositions of the control and treatment groups provide more opportunities for differential effects in how respondents relate to the instrument – that is, measurement invariance may be more plausible in a DiD design than, say, an RCT design. For example, if there are more poor students in the treatment group that can

be adjusted for statistically, but if poverty affects how students respond to a survey instrument that impact on measurement will not be accounted for. If a math engagement instrument is particularly inaccurate for poor students because they are more likely not to complete surveys and the treatment improves poor student's math engagement greatly the treatment estimate would be biased down. This problem might be even worse for DiD designs that have differential timing in treatment where multiple units experience treatment at different times, and there may be secondary complications around the evidence supporting the parallel trends assumption.

This paper proceeds as follows. In section 1, I provide background that explains how using multi-item instruments to measure an outcome can affect DiD designs, and the contexts and conditions under the associated problems that are most likely to be a problem in education DiD studies; in section 2, I discuss the design and analysis plan of a Monte Carlo simulation for evaluating the bias caused by scoring decisions for multi-item instruments ; in section 3, I present the results of the Monte Carlo simulation study; and in section 4, I discuss implications of results and offer recommendations for education researchers using the DiD design with multi-item instrument outcome data.

Background

The DiD Method

First used in the 1850's by John Snow in his analysis of the cause of cholera (Snow, 1855), DiD is one of the oldest quasi-experimental research designs. Its logic revolves around comparing the different trajectories of entities that are affected by a treatment, and other entities that are not affected by the treatment (Dimick & Ryan, 2014). Usually, these entities are grouped in cross-sectional units such as schools or districts. Both the time period (denoted by t) and group of an observation of an entity (denoted by g) is crucial to obtaining a causal estimate. This is

because it allows for the generation of potential outcomes. Concretely, let D_{gt} represent an observation at time *t* and in group *g*. $D_{gt} = 1$ if a unit is exposed to treatment at time *t,* otherwise, $D_{gt} = 0$. To obtain a causal estimate we need the outcome for both when that observation is treated ($Y(1)_{gt}$) and when the observation is not treated ($Y(0)_{gt}$). Given that we cannot observe both, in the DiD design we use untreated observations to generate $Y(0)$ for treated observations based on the assumption that important unmeasured variables are either time-invariant group attributes or time-varying factors that are group invariant, the parallel trends assumption (Wing et al., 2018) . This is done through the equation:

$$Y(0)_{gt} = a_g + b_t + \varepsilon_{gt}$$

(1)

where a_g represents combined effects of time-invariant characteristics of group *g* and b_t represents the combined effects of the time-varying but group-invariant factors. Integrating the observed outcomes with this equation leads to the generalized DID estimating equation:

$$Y_{gt} = a_g + b_t + \delta D_{gt} + \varepsilon_{gt}$$

(2)

To highlight the possible effects of measurement we simplify from this generalized form to a situation where there are only two groups, one treated and one not, and two time points, pre and post treatment. This is the simplest version of the DiD design known as the 2X2 (two by two) DiD. With only two time periods and two groups we can represent group membership and the time period using dummy variables, T_g and P_t respectively. This means the treatment effect is an interaction between the two ($D_{gt} = T_g$ x P_t). Thus, the estimating equation for the 2X2 DiD is:

$$Y_{gt} = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \text{ x } P_t) + \varepsilon_{gt}$$

(3)

In the above, β_1 represents the unit difference between the two groups before the treatment on the outcome, β_2 represents the unit change in the outcome for the control group, and the interaction term, β_3, represents the mean treatment effect (the parameter of interest). Notably Y_{gt} is typically treated as a single variable that is measured perfectly.

Imperfect Measurement and the DiD Design

Treating the outcome in this way introduces an estimation assumption that is separate from the identification assumptions of the DiD design. This assumption of perfect measurement is unlikely to be met for most educational constructs, especially those measured by surveys. Any given item on a survey is an imperfect representation of the target latent construct and so most surveys use multi-item instruments to obtain a better measure. When using such a multi-item instrument, Y_{gt} is a composite of several items rather than a single variable. Accordingly, one problem that can arise is if the items are scored in such a way that items are mis-weighted. Scoring approaches can be as simple as taking the sum of every item in the instrument (a so-called "sum score") or complex as using measurement models based on Item Response Theory (Van der Linden & Hambleton, 2013). Regardless of the method used, the computed score is an imperfect representation of the outcome, which is often a latent variable (e.g., student achievement or self-efficacy). If the items are substantially mis-weighted this can result in a biased causal estimate. To show how this can work, let us begin with a case where there is only a single latent variable (our dependent variable) of interest, and that latent variable is regressed

on indicators of time, treatment, and treatment by time. A possible measurement model for this latent variable using four items for person *i* is portrayed in Figure 1.

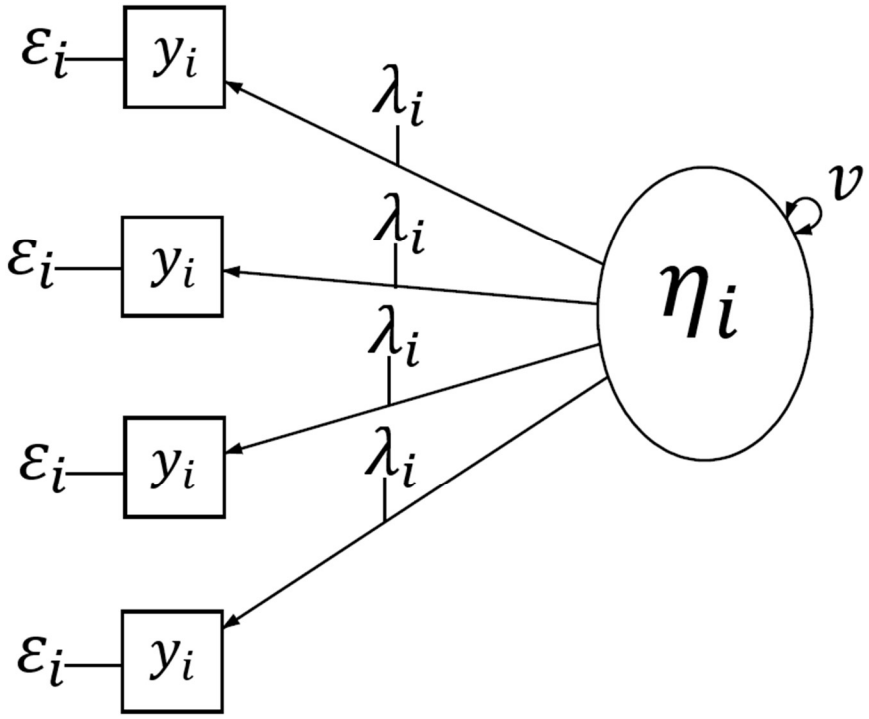

Figure 1. *Example individual measurement model*

To further simplify we focus on a single item that we assume is continuous. While the potential for mis-weighting is magnified when there are multiple items, mis-weighting effects can be demonstrated in a single item. For this single item case, in equation form, the measurement model can be translated as:

$$y_i = v + \lambda\eta_i + \varepsilon_i \quad (4)$$

Where $\boldsymbol{y}_i$ is the observed item response for item *j*, $\boldsymbol{v}$ is the intercept, $\boldsymbol{\lambda}$ is the loading for that item, η_i is the single latent outcome for person *i*, and ε_i is the residual (with a mean of zero) where $VAR(\varepsilon_i) = \boldsymbol{\theta}$.

We can then use a slight variation of equation(3) in which the latent outcome rather than the score is the dependent variable, to provide the framework to use the measurement model to produce a causal estimate.

$$\eta_i = \beta_0 + \beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \text{ x } P_t) + \varepsilon_{gt}$$

(5)

As the new equation's dependent variable is now latent it needs to be put back in terms of the observable item responses. This can be done by plugging equation (5 into the measurement model given by equation (4 assuming $\boldsymbol{\nu}$ is 0 (done for convenience, though not necessary) and β_0 is 0 (i.e. the control group average is 0). This gives us :

$$y_i = \lambda(\beta_1 T_g + \beta_2 P_t + \beta_3 (T_g \text{ x } P_t) + \varepsilon_{gt}) + \varepsilon_i$$

(6)

We can then determine what the expected value is to derive the causal estimate.

Given $E(\epsilon_i)$ = 0 and $E(\varepsilon_i)$= 0,

$$E(y_i) = E(\lambda(\beta_1 T_g + \beta_2 P_t + \beta_3 T_g P_t))$$

(7)

For the control group before treatment is applied, $a_i = 0$ and $t_i = 0$ so the expectation of the observed score for $E(y_i)$ is zero. For the control group after treatment is applied $a_i = 0$ and $t_i = 1$ so

$$E(y_i \mid T_g = 0 \ \& \ P_t = 1) = \lambda \beta_2 E(t_i) = \lambda \beta_2$$

(8)

For the treatment group before treatment the expectation is:

$$E(y_i \mid T_g = 1 \,\&\, P_t = 0) = \lambda \beta_1 E(t_i) = \lambda \beta_1$$

(9)

Finally for the treatment group post-treatment no terms have an expectation of zero so:

$$E(y_i \mid T_g = 1 \,\&\, P_t = 1) = \lambda \beta_1 E(t_i) + \lambda \beta_2 E(t_i) + \lambda \beta_3 E(t_i) = \lambda \beta_1 + \lambda \beta_2 + \lambda \beta_3$$

(10)

Therefore, the expectation of the observed scores between control and treatment groups adjusted for baseline differences is:

$$[E(y_i \mid T_g = 1 \,\&\, P_t = 1) - E(y_i \mid T_g = 1 \,\&\, P_t = 0)] - [E(y_i \mid T_g = 0 \,\&\, P_t = 1) -$$

$$E(y_i \mid T_g = 0 \,\&\, P_t = 0) = (\lambda \beta_1 + \lambda \beta_2 + \lambda \beta_3) - (\lambda \beta_1) - (\lambda \beta_2) - 0 = \lambda \beta_3$$

(11)

As Equation(11 shows, the difference in the means of the control and treatment observed scores adjusted for differences in the observed score at baseline would be the true treatment effect, β_3, weighted by the item loading (λ). This means if you use the observed score and the item loading is less than one the estimated treatment effect will be lower than the true treatment effect and if the item loading is greater than one the estimated treatment effect will be greater than the true treatment effect. For example, imagine a treatment results in a one unit increase on one item. If the loading of that item on the latent variable is .4 a one unit increase is really a .4 increase on the outcome you are interested in. Thus, mis-specifying the loadings can bias a treatment effect even in a simplified situation where the assumptions of the DiD are met and it mimics an experiment perfectly.

While we demonstrated this problem on a single item, the problem emerges similarly for multiple items. Instead of the causal estimate being weighted by a single loading, it is weighted by a vector of the loadings of the component items. A common approach like sum scoring that forces the loadings to be equal across items when they are not could bias the estimate of a treatment's effect on the treated.

For example, the Gates Foundation's Measures of Effective Teaching project ("Learning about Teaching," 2010) used a survey scale of seven Likert items to measure student effort. A common way to transform the seven questions that make up the scale into a score for analyses is to add up the item responses and divide by the number of items to produce a sum score. As McNeish and Wolf (2020) demonstrate, this practice translates to a measurement model in which each item is contributing equally to the construct and the residual variances are constrained to be equal. In terms of equation 3, this means for each item $\boldsymbol{\lambda}$ and ε_i are the same constants. Depending on the scale this might be an unlikely assumption. For example, Beck's Depression inventory, a common scale used to measure depression (Beck & Beamesderfer, 1974), has items related to sleep patterns and subjective emotional experience making it unlikely that it is appropriate to treat those items as equivalent. Thus, bias can be introduced by using sum scores if not all the items should have equal weights (i.e., there are not equal associations between the latent variable and the observed item responses), but that constraint is imposed anyway.

Bias induced by mis-weighting is not unique to DiD designs. Prior work has shown how the same problem manifests itself in experiments and regression discontinuity designs (Soland, 2022; Soland et al., 2022). However, in the DiD design this problem entails additional complications because the use of non-equivalent groups and the resultant reliance on the parallel trends assumption opens up the possibility that there will be measurement invariance between

the groups and timepoints and that this will violate the parallel trends assumption. This means that each group at each timepoint must be treated separately as in Figure 2, which introduces more potential for bias, as loadings and intercepts do not cancel out (See Appendix I for supporting derivations).

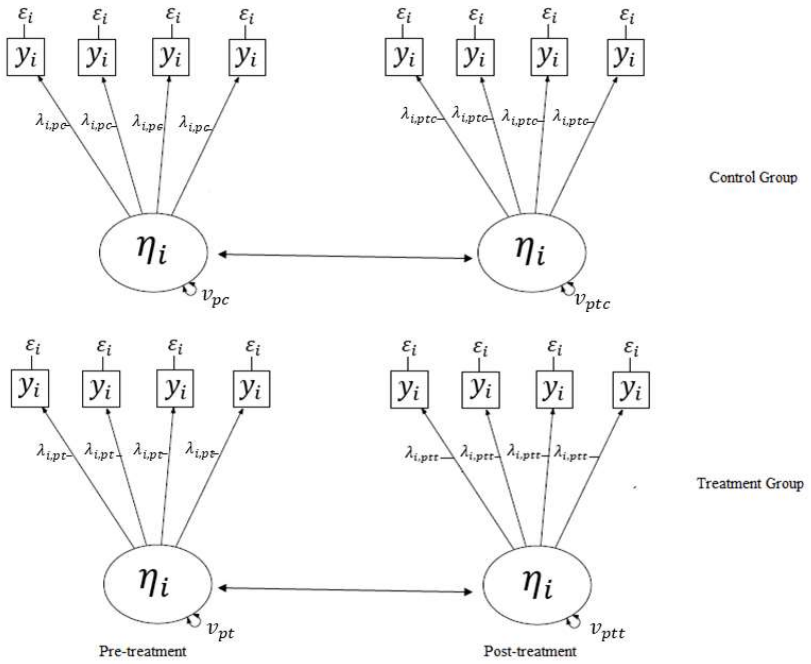

Figure 2. *Example DiD measurement model*

Plausible Measurement Scenarios

How this bias manifests exactly depends on which time points and groups have measurement noninvariance. In this paper we investigate three different measurement noninvariance scenarios. In the first scenario there is no measurement noninvariance or in other words the model is measurement invariant. Accordingly, there is no violation of any identification

assumptions and there should be no impacts on the causal estimate or evidence of parallel trends unless there is mis-weighting.

In the second scenario there is measurement noninvariance between the control and treatment groups, with differences in the two groups' loadings. Differences in loadings might occur due to the different compositions of the control and treatment groups. For example, the control group might have more students that refuse to answer one item leading to a weaker loading on the latent outcome.

In this scenario even though the identification assumptions are met the typical estimating equation does not account for different loadings across groups, which can result in bias in the causal estimate. In addition to these direct effects on the causal estimate, this form of measurement noninvariance has the potential to influence the evidence needed to support the parallel trends assumption. Pre-treatment trends are often examined both visually and through statistical tests (Angrist & Pischke, 2008; Cunningham, 2020) to establish the plausibility of the parallel trends assumption. Biased scores have the potential to distort pre-treatment trends affecting both the perceived credibility of the parallel trends assumption and any decisions around compensating for violations of the parallel trends assumption. Importantly, this is distinct from the validity of the parallel trends assumption itself.

In the third scenario, the form of measurement noninvariance the treatment group at the post-treatment timepoint is measurement non-invariant with the other three groups. This form of measurement noninvariance is usually the result of the treatment changing a respondent's internal standards of measurement, values, or conception of the measured construct and can be categorized into three types of response shifts (Sprangers & Schwartz, 1999). Each type is manifested through changes in item loadings, thresholds, or intercepts. Recalibration, a change

in the respondent's internal standards of measurement corresponds to a change in threshold or intercept values. Reprioritization, a change in the importance a respondent places on an item when thinking about the outcome, is manifested by an increase or decrease of item loadings. Finally, reconceptualization, a change in the respondent's understanding of the target construct, results in a reduction of one or more item loadings to zero or a new loading(s) onto new item(s).

Response shifts are a direct violation of the parallel trends assumption because they are neither a time-invariant group attribute nor a time-varying factor but still affect the measured treatment. Thus, they can be expected to bias the causal estimate directly as well as affect the evidence for parallel trends. However, unlike other violations of the parallel trends assumption this violation, can be accounted for by letting the loadings vary. To what extent each form of measurement noninvariance and the degree of measurement noninvariance affects analysis of a DiD design given a researcher's scoring strategy is a key question yet to be explored.

The Current Study

The goal of the current study is to investigate how measurement model misspecification affects analysis of a DiD design in the context of the current dominant scoring strategy of using sum scores. I am examining three forms of measurement model misspecification and their impact on bias in DiD estimates: improper use of sum scores, noninvariance between control and treatment groups, and response shifts.

Three research questions will be addressed in three simulation studies.

1. To what extent does model misspecification due to misapplication of sum scoring in a DiD design affect bias in the produced estimates?
2. To what extent does model misspecification due to measurement noninvariance between groups in a DiD design affect bias in the produced estimates?

3. To what extent does model misspecification due to a response shift in a DiD design affect bias and statistical significance in the produced estimates?

Answering these three questions could help establish a clear picture of the potential problems and benefits of the various strategies to combining multiple items so that they can be used in the DiD analytic approach. This would help researchers make informed decisions about what strategy is appropriate for a given situation.

Simulation Specifications

Overall Data-generating Process

To answer the research questions, I produced three variations of a shared data generating process, one for each simulation study. For the first study, measurement invariance holds in the data-generating model, and the factor loadings of the items are varied. For the second, there was measurement noninvariance between the treatment and control group in the data-generating model. And for the third, a response shift results in measurement noninvariance in the post-treatment treatment group only.

Besides these variations the simulated data for each study was generated using the same data-generating model. Figure 3 below shows this data-generating model, which mimics a scenario where two groups are slowly increasing on a measure, but one group is lower than the other by a set amount until there is a plateau. In Figure 3, y_i represents *i* item, λ_i represents the loading for y_i, $\mu(\eta_n)$ represents the mean of a latent variable at *n* timepoint, σ_n^2 represents the variance of a latent variable at *n* timepoint, and t represents the treatment.

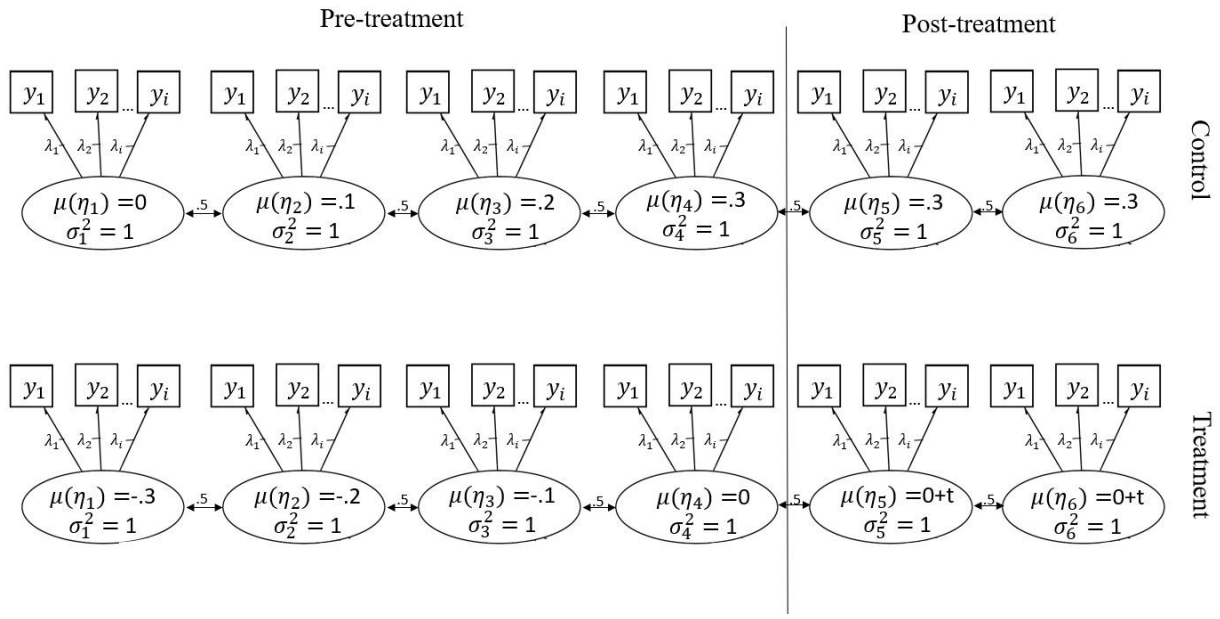

Figure 3. *Data-generating Model*

This model was generated with Mplus version 8.7 (Muthén & Muthén, 2017) in conjunction with the R package Mplus automation (Hallquist & Wiley, 2018) and assumed that the true generating model for the outcome was a latent variable that explained responses to a set of observed items. The items were generated as continuous variables for simplicity and to generalize to many different types of multi-item outcomes. Six timepoints were generated: four pre-treatment and two post-treatment with correlations of .5 between the latent variables at each timepoint to mimic the mean .46 autocorrelation found by Barnard-Brak et al. (2021) . These six timepoints were generated for two groups, a group that did not receive treatment, the "control group", and a group that did, the "treatment group." For both groups, the latent means were standardized to the first time point of the control group. The control group had a latent mean of 0 SDs at timepoint 1, .1 SDs at timepoint 2, .2 SDs at timepoint 3, and .3 SDs at timepoints 4-6. In contrast, the treatment group had a latent mean of -.3 SDs at timepoint 1, -.2 SDs at timepoint 2,

-.1 SDs at timepoint 3, 0 SDs at timepoints 4, and means equal to the treatment effect at timepoint 5 and 6. The variance for the latent variables at each timepoint for both groups was set to 1, in part to facilitate interpretation of the loadings as fully standardized.

Systematically Varied Conditions in the Data-generating Process

From this base-data generating model several conditions were varied. The key condition that separated the three simulation studies was the presence and degree of measurement noninvariance. In addition to this condition, selected other conditions were varied to represent a wide range of realistic scenarios.

Presence and Degree of Measurement Noninvariance

The primary factor that was varied was the presence of measurement invariance. In the first simulation study, addressing research question 1, measurement invariance was established across timepoints and groups which meant the factor loadings were the same at every timepoint (though they differed by item). In the second simulation study, addressing research question 2, there was measurement invariance between the control and treatment groups, which meant there were differences in the loadings for the same item at the same timepoint between the control and treatment groups, but within each group the parameters were invariant. Finally, in the third simulation study, addressing research question 3, there was measurement noninvariance in the treatment group post-treatment (timepoints 5 and 6 in the treatment group), but the other item parameters were invariant across groups and over time.

Within the second and third simulation studies the degree of measurement invariance was also varied. There is little direct research on the plausible degree of measurement noninvariance between the control and treatment groups in a DiD design, and it is likely dependent on how similar the groups are. Accordingly, for the second simulation study I evaluated a small and

medium magnitude of measurement noninvariance in the loadings between the control and treatment groups. This meant that in the small magnitude condition there was a positive increase of 0.10 in the loadings of items 1 & 4 and in the medium magnitude of noninvariance there was a positive increase of 0.25 in the loadings of items 1 & 4. These values were chosen based on previous research that defined these differences when the factor was standardized for one group (Stark et al., 2006; Yoon & Millsap, 2007). For both the 4-item and 8-item condition these two items were the sources of measurement noninvariance. This means that the proportion of invariant items for the 8-item condition was smaller than that of the 4-item condition. The fixed number of invariant items reflects a plausible scenario, where a larger scale that contains invariant items is reduced to save money or time and the shorter form of the scale contains the invariant items.

For the third simulation study, the degree of measurement invariance was determined by the type of response shift. I tested two of the response shifts outlined by Schwartz & Sprangers (2000; 1999). Reprioritization was simulated through the increase of item 4's loading by .4, reflecting the magnitude observed by Fokemma et al. (2013). This corresponds to respondents finding item 4 more relevant to the measured construct in the treatment group after intervention. This type of response shift might be caused by the treatment increasing a respondent's understanding of an item or triggering the recall of relevant information. To simulate reconceptualization, a change in the understanding of a construct, item 4's loading was reduced to zero. This type of response shift occurs if the treatment changes respondents' understanding of a construct enough to render an item irrelevant or make an additional item relevant. For example, this might occur if there is a common misconception about the construct that treatment changes.

Factor Loadings

Within each simulation study, the most important condition that was varied was the pattern of factor loadings. Three plausible factor loading patterns were chosen: a high condition where all standardized item loadings fell between .7-.9, a low condition where the item loadings fell between .3-.5, and a mixed condition where item loadings fell between .1-.8. Item loadings can be squared to obtain the proportion of the variance in the item that the latent variable explains. Accordingly, the high condition reflected a well-designed scale in which the latent variable explains a high amount of variance for each of the items, the low condition reflected a poorly designed scale in which the latent variable explains a low amount of variance for each of the items, and the mixed condition represented a scale in which the latent variable explains a lot of variation for some items and little for others.

Other Conditions Varied During Data Generation

In addition to measurement invariance and factor loadings, three other conditions were varied. The first was sample size. Two sample sizes were chosen: 500 and 3700 per group. These values were chosen to reflect the average number of students of two commonly used groups in education, public schools and districts (Snyder et al., 2019). The second was number of items comprising the outcome variable. 4 and 8 were identified as representative of the length of many social-emotional survey scales (Wang et al., 2016; West et al., 2018), as well as subscales of observation protocols (Goodman, 1997; Pianta et al., 2008). The third was treatment effect size. The majority of educational interventions that are well powered resulted in effect sizes less than .2 (Cheung & Slavin, 2016; Kraft, 2020; Rocconi & Gonyea, 2018).

Accordingly, .1 and .2 were chosen as plausible treatment effect sizes and a null effect was tested for comparison.

For each group of datasets generated, every condition, summarized in Table 1, was crossed with every other condition within each simulation study. For every set of unique conditions 100 datasets were produced. Full data generating code is available at https://osf.io/jp2xs/.

Table 1. Summary of Conditions of the Data-generating Process

Condition	Options	Simulation Studies
Item Loadings	High, Low, Mixed	1, 2, 3
Sample Size	500 & 3700	1, 2, 3
No. of Items	4 & 8	1, 2, 3
Treatment Effect Size	0, .1, .2, ,3	1, 2, 3
Control-Treatment Noninvariance	Small (.1 increase), Medium (.25 increase)	2
Response Shift	Reconceptualization, Reprioritization	3

Analyzing Generated Item-level data

For each simulation study two types of analysis were used. The first was the commonly used approach of averaging items to produce a score and then using parametric regression to produce a causal estimate. The second used structural equation modeling (SEM) to model each item's relationship with the latent construct by group and timepoint. For the second and third simulation

studies both a correctly specified SEM model and an incorrectly mis-specified model that assumed measurement invariance was used. Below I detail the SEM models further.

SEM Models

For every SEM analysis, I used a structured mean model with multiple groups to model the DiD data. In this model each timepoint in the control and treatment groups were treated separately, which allows for tests of measurement invariance between timepoints in groups using the procedures laid out in (Oort, 2005). Rather than only focusing on a single causal estimate, bias can be examined in the latent means and variances by timepoint for each group such that the effects on differences in trends can be better dissected.

The three variations of this model that were used to in the simulation studies are shown in Figure 4 below.

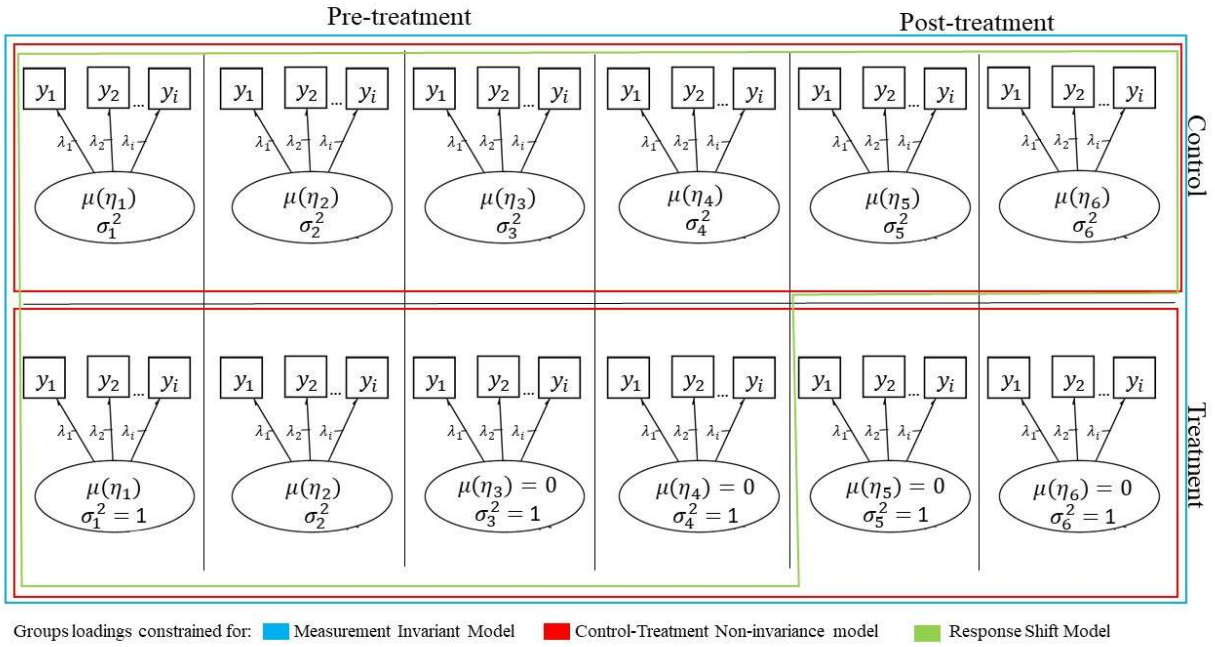

Figure 4. *Loading Constraints for Three Measurement Models*

The measurement invariant model (SEM (MI)) constrained the loadings of each item between every timepoint in the control and treatment groups to be equivalent. This model was used in all three simulations studies. For simulation studies 2 and 3 the presence of noninvariance meant it was a mis-specified model for those studies. The Control-Treatment Noninvariance model (SEM (CT)) constrained loadings of each item between every timepoint within the control and treatment groups to be equivalent but allowed them to vary between the two groups. This model was only used for simulation study 2, as the correctly specified model. Finally, the response shift model (SEM (RS)) constrained all groups and timepoints loadings except the treatment group post-treatment to be equivalent. This model was only used for simulation study 3, as the correctly specified model. Table 2 summarizes the type of analysis used for each simulation study.

Table 2. Summary of Analysis Models by Simulation Study

	Sum Scoring	SEM (MI)	SEM (CT)	SEM (RS)
Simulation Study 1	X	X		
Simulation Study 2	X	X	X	
Simulation Study 3	X	X		X

While modeling latent variables directly is preferred because latent variables are difficult to capture through scoring (Grice, 2001), the SEM models described above can be used to create factor scores that can be used in conjunction with parametric models, which may be useful for researchers who want to improve measurement but use tools developed outside a SEM

framework. Results included in Appendix II demonstrate that factor scores produce treatment estimates that are close to those produced by modeling the latent variables directly.

Evaluation Criteria

The primary focus of this study was on the estimate of means by timepoint and group which was key to both the treatment effect and the establishment of parallel trends. I examined the bias and variance in the means, as well as the bias in the treatment estimate for each analytic method. In addition, to these two aspects I also evaluated the recovery of the loading parameters as a check on how well each measurement model recovered the parameters of the data generating process.

Results

Across all studies, there were no convergence failures, and parameter recovery when fitting the true model to the generated data was excellent (see Appendix II Table 6).

Simulation Study 1: Impact of Misapplication of Sum Scoring

In datasets in which measurement invariance held across groups and timepoints sum scores induced bias that ranged from -.076 to .009. As seen in figure 4 there were clear patterns. Sum scores performed best when there were more items, and the loadings were high. The amount of bias was proportional to the size of the treatment effect with more bias induced for larger effects because the mis-weighting of sum scores has a multiplicative impact. Interestingly, in the 4-item condition larger sample sizes increased bias. This is likely because the larger sample amplified

the effect of the mis-weighting. In contrast, there was very minimal bias across conditions and treatment effects when using the SEM model ranging from -.003 to .009.

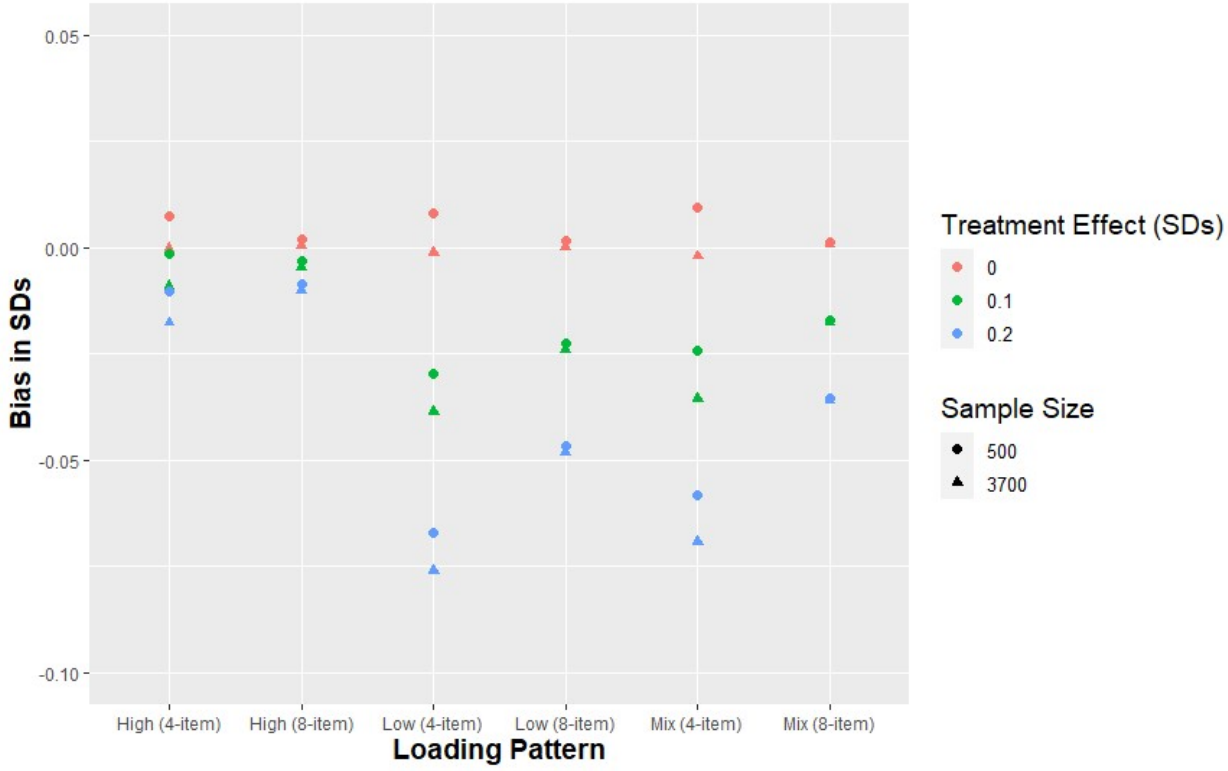

Figure 5. *Sum score bias by loading pattern for Simulations Study 1*

Examining the mean estimates underlying the treatment effect estimates, reveals a more comprehensive picture of the impact of analysis type. Consider the condition in which the treatment effect estimate is most biased, the low loading, 4-item, large sample size condition, shown in Figure 6 below. Sum scoring biases the estimated means, so that the estimated pre-treatment trends are not parallel. While the true slope of both the control and treatment group is .1, the median estimated slope is .064 for the control group and .062 for the treatment group. In contrast, SEM recovers the means of each timepoint accurately and recovers the true pre-treatment trends. However, there is a tradeoff with variance because the estimation of SEM

loadings entails greater variance in the estimated latent means. The SEM mean estimates standard deviations were twice as large as that of sum scoring standard deviations (.03 vs. 015) for this condition.

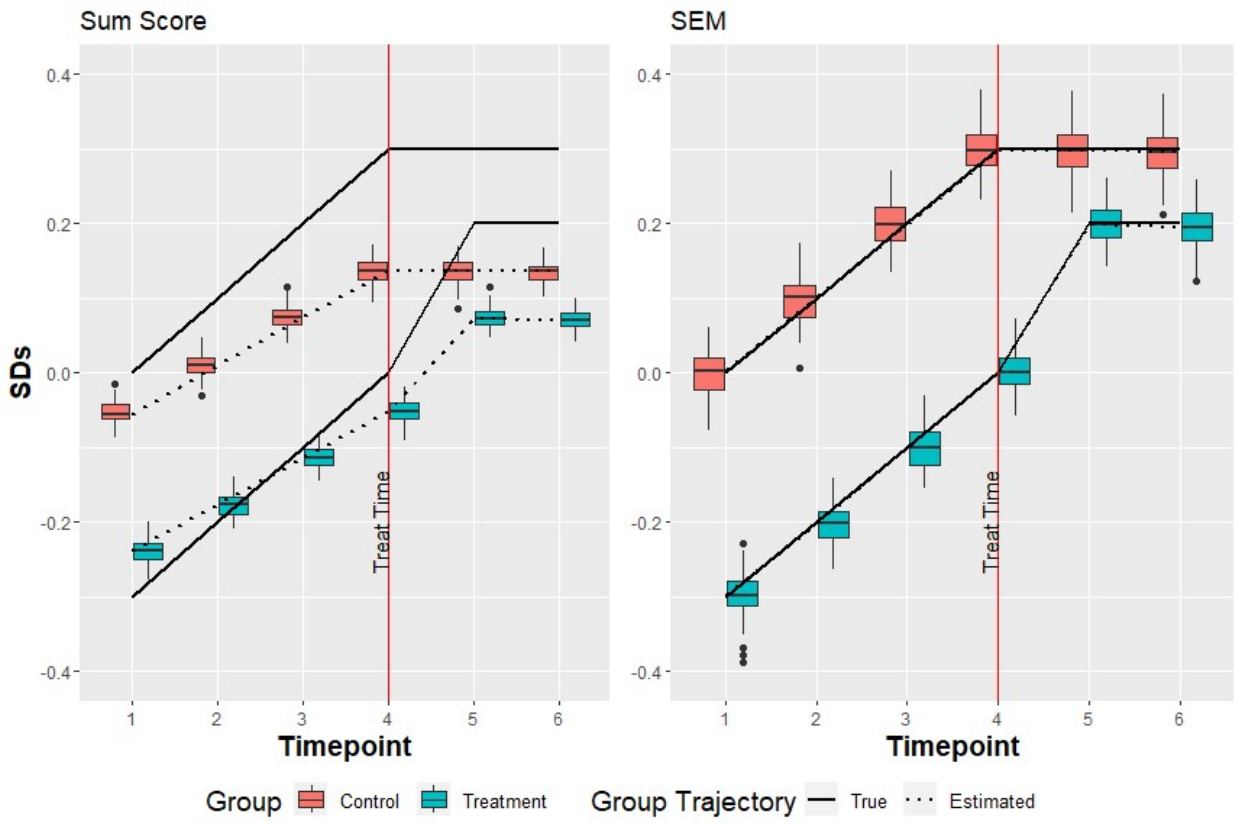

Figure 6. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, Low Loading, N=3700, 4-item condition in Simulation Study 1*

Even when conditions are ideal for sum scoring and the treatment estimate is minimally biased sum scoring still has biased mean estimates. Figure 7 shows the means for the high loading, 8-item, small sample size condition. In this condition the individual means are still biased but they are biased by similar amounts in each group allowing for the minimally biased recovery of the treatment effect and the recovery of parallel pre-trends. Additionally, there is a proportionally smaller variance trade-off (.04 SD vs. 05 SD) between sum scoring and SEM.

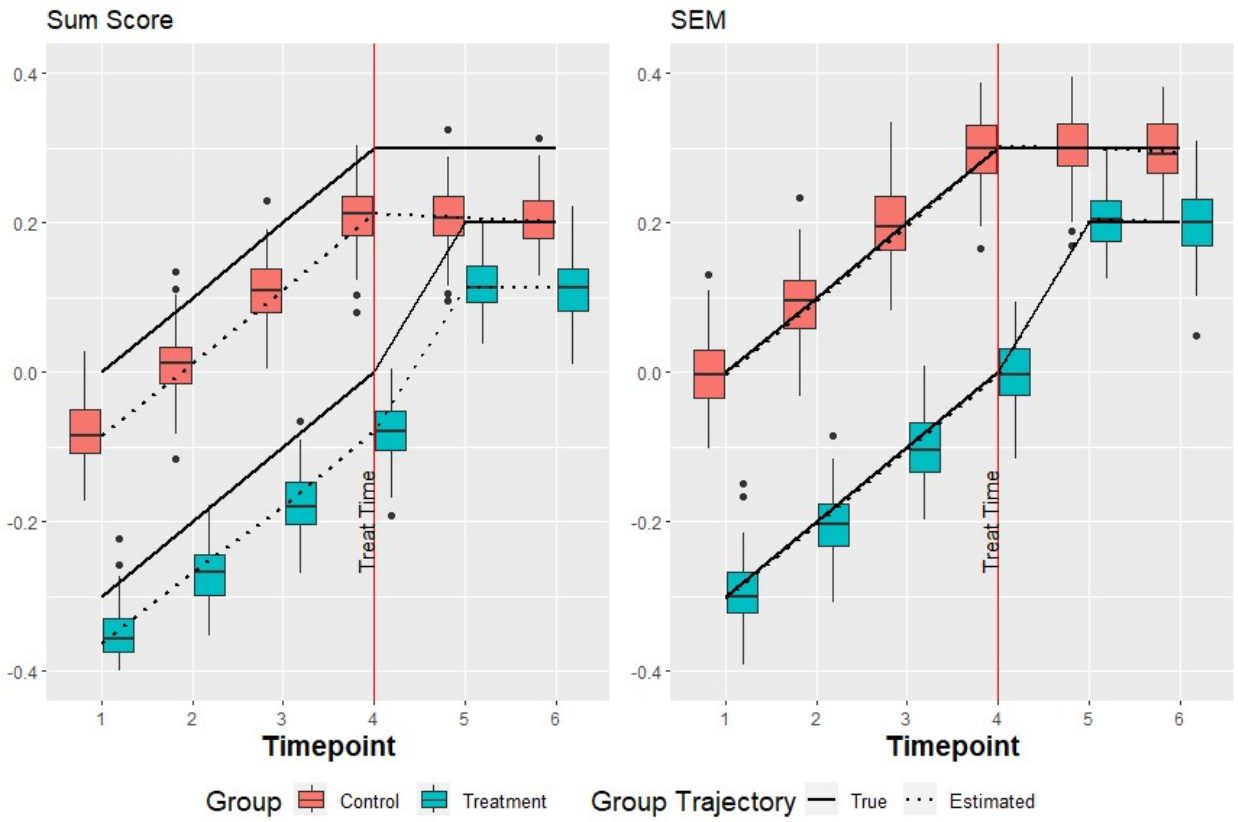

Figure 7. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, High Loading, N=500, 8-item condition in Simulation Study 1*

Simulation Study 2: Impact of Control-Treatment Measurement Noninvariance

For the datasets that had Control-Treatment noninvariance each scoring strategy induced some bias. This was extremely small in the case of SEM (CT) (.001 to .012) because the model matched the data-generating process, but was larger in sum scoring and for SEM (MI). As shown in Figures 8 the amount of bias was affected by the amount of CT noninvariance with a greater amount of CT noninvariance shifting the bias of every scoring strategy up. This meant that sum scoring produced treatment effect estimates that were closer to the true treatment effect when there was a medium amount of CT noninvariance but increased the amount of bias for SEM

(MI). This pattern of bias is probably due to the increase in loadings in the treatment group vs. the control group. The reverse would likely result in negative bias that would exacerbate the negative bias seen with sum scores in simulation study 1. Like sum scoring the amount of bias in the SEM (MI) treatment estimates was proportional to the treatment effect estimate size.

Figure 8. *Bias by loading pattern, .2 treatment effect for Simulation Study 2*

Biases in the treatment effect estimates are reflected in the means by timepoints as shown in Figure 9 below. In the low loading, 4-item, large sample size condition shift, CT noninvariance results in lower mean estimates for timepoints 5 and 6 in the control group, which results in a larger treatment effect estimates. It also distorts recovery of parallel trends further. In the presence of medium CT noninvariance the sum score produced median estimated slope is .06 for

the control group and .07 for the treatment group and the SEM (MI) produced median estimated slope is .085 for the control group and .113 for the treatment group. The variance tradeoff remains the same as it is in Simulation Study 1.

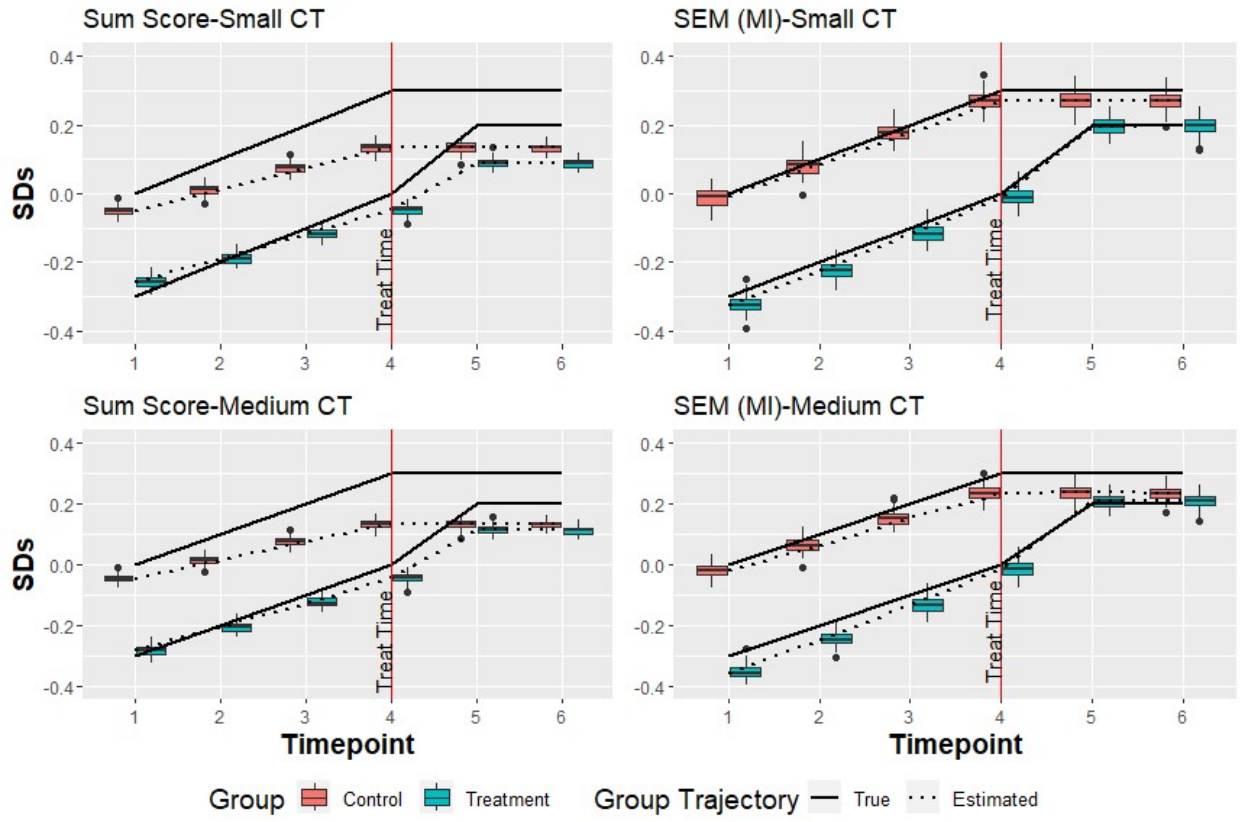

Figure 9. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, Low Loading, N=3700, 4-item condition in Simulation Study 2*

Similar to simulation study 1, when conditions are optimal for sum scoring, such as the high loading, 8-item, small sample size condition sum scoring still induces some bias in the mean estimates but this bias is mostly consistent and only shifts all the mean estimates down. There is only a negligible impact on SEM (MI).

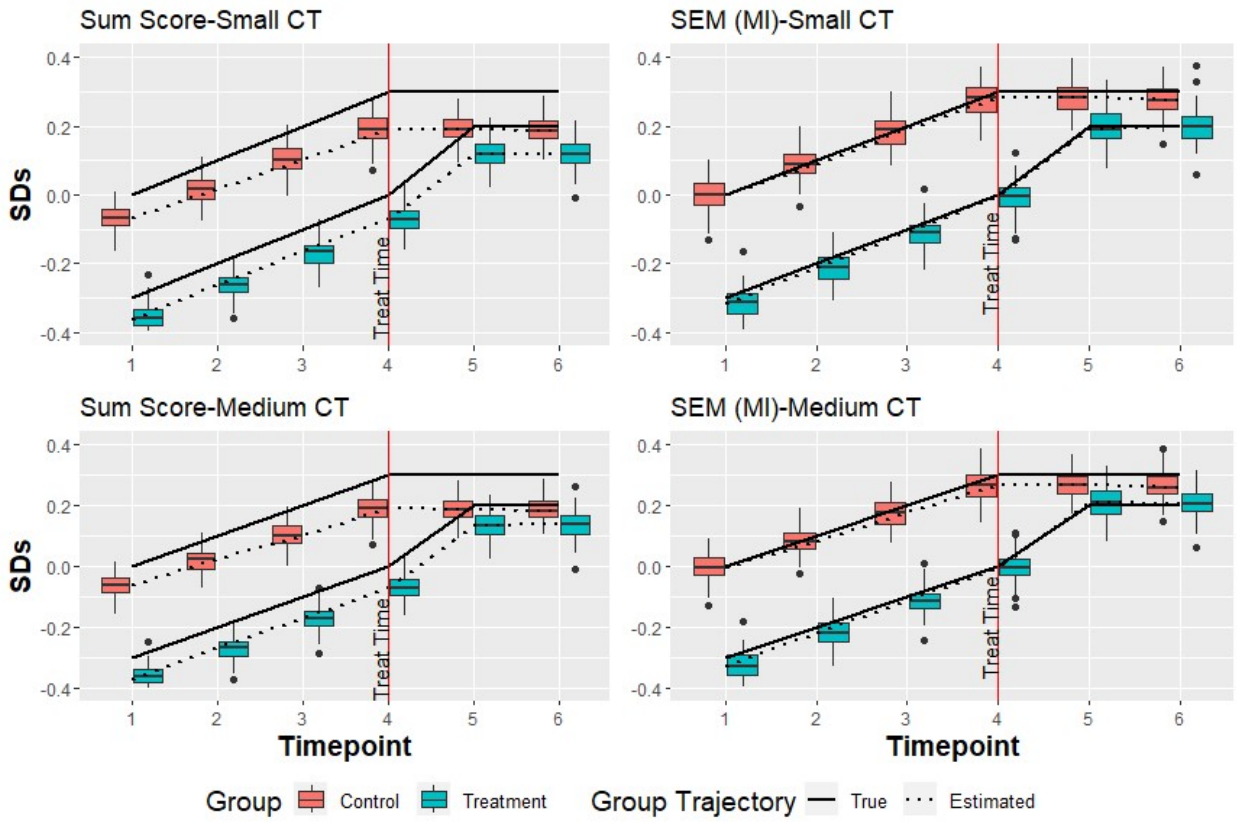

Figure 10. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, High Loading, N=500, 8-item condition in Simulation Study 2*

Simulations Study 3: Impact of Response Shifts

Like in the other two simulations studies SEM that allowed for noninvariance (SEM (RS)) had the least amount of bias. However, for both reconceptualization and reprioritization, SEM (RS) had notable bias for the 4-item mixed loading conditions that ranged from -.034 to .01 but had less bias (0 to .01) for the other conditions. For SEM (MI) and sum scoring the amount of bias induced was dependent on loadings, sample size, and number of items as shown in Figure 11. These patterns were in line with the patterns in the other simulation studies except the 4-item mixed loading condition was notably biased.

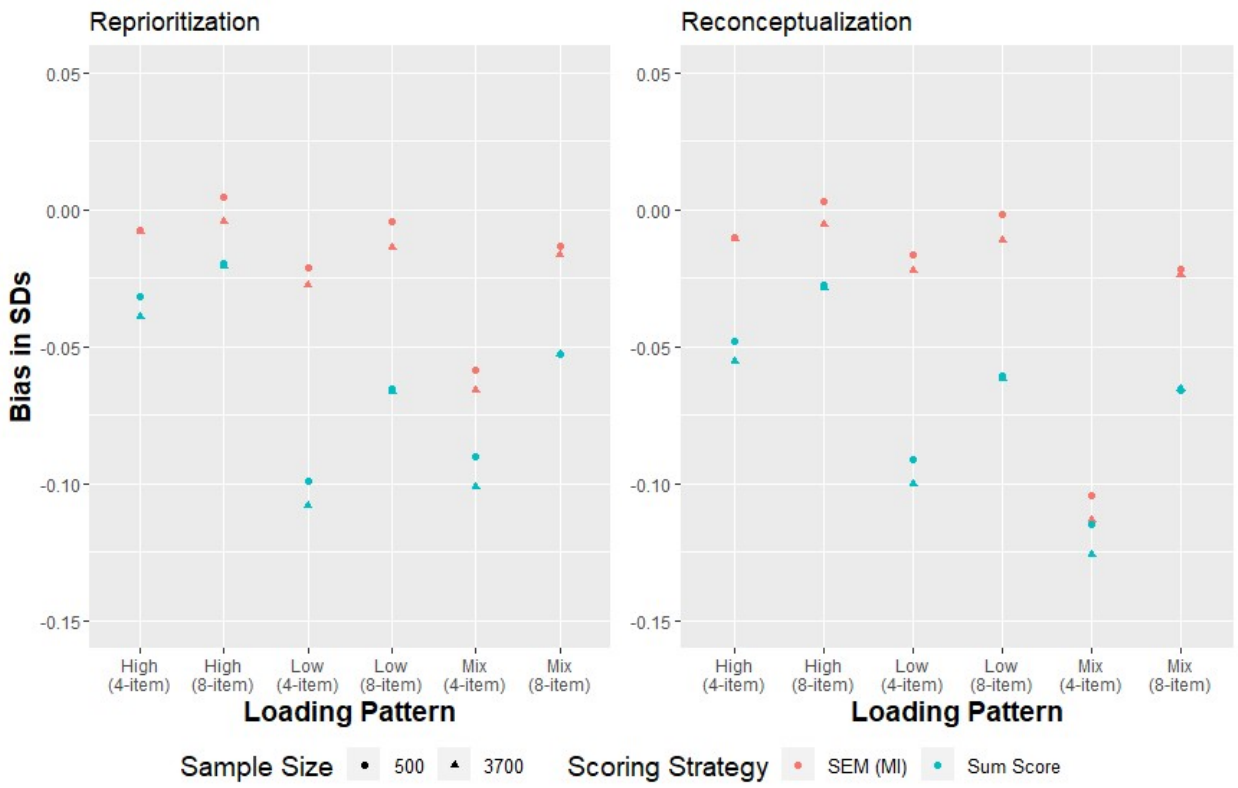

Figure 11. *Bias by Loading Pattern, .2 Treatment Effect for Simulation Study 3*

Examining the 4-item mixed condition more closely as shown in Figure 12 below it is clear that the treatment effect bias is the result of smaller mean estimates in the treatment group at timepoints 5 and 6. Unlike CT noninvariance, neither type of response shift prevented the recovery of parallel pre-treatment trends. The variance tradeoff was particularly high (.015 SD vs. .055 SDs) likely because of the increased variability in the loadings caused by the response shifts.

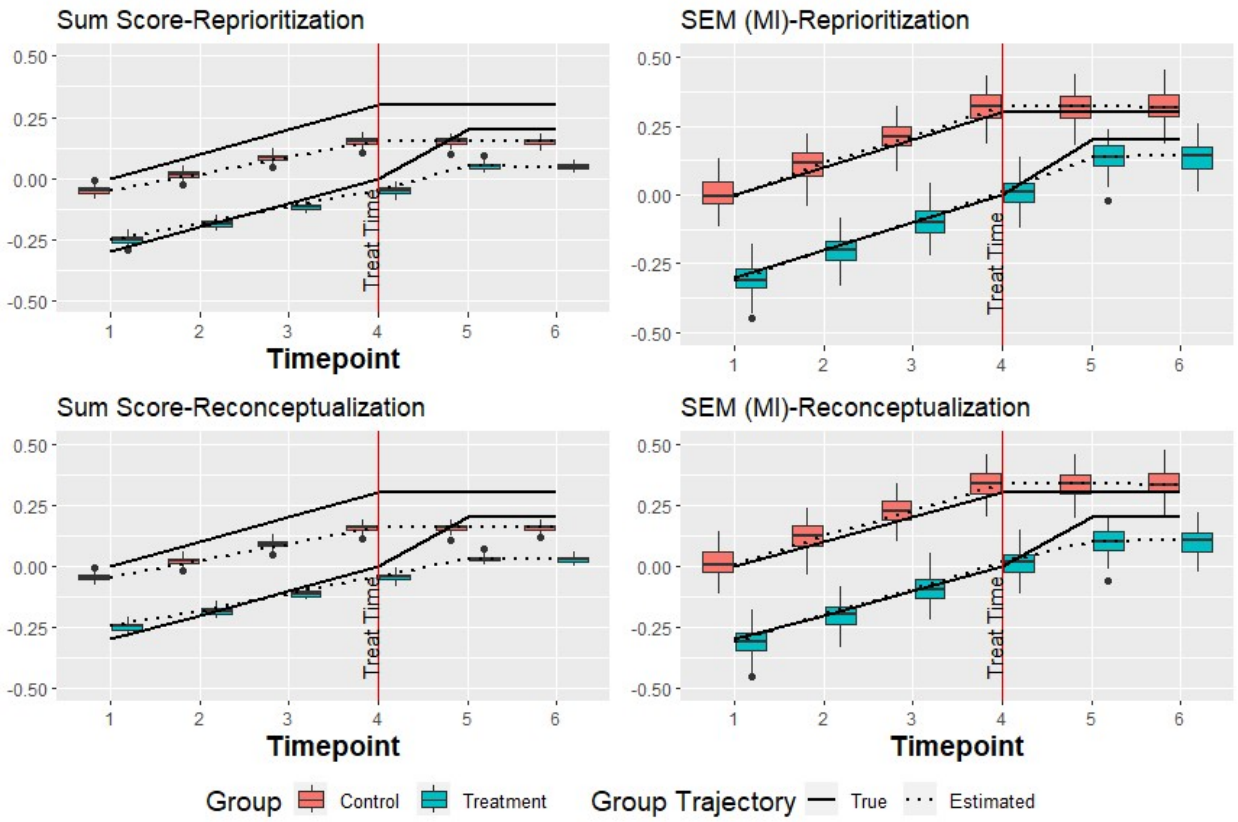

Figure 12. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, Mix Loading, N=3700, 4-item condition in Simulation Study 3*

When the conditions were more optimal for sum scoring, like in the high loading, 8-item, small sample size condition shown in Figure 13, there were clear deviations from the true mean estimates that led to an underestimation of the treatment effect.

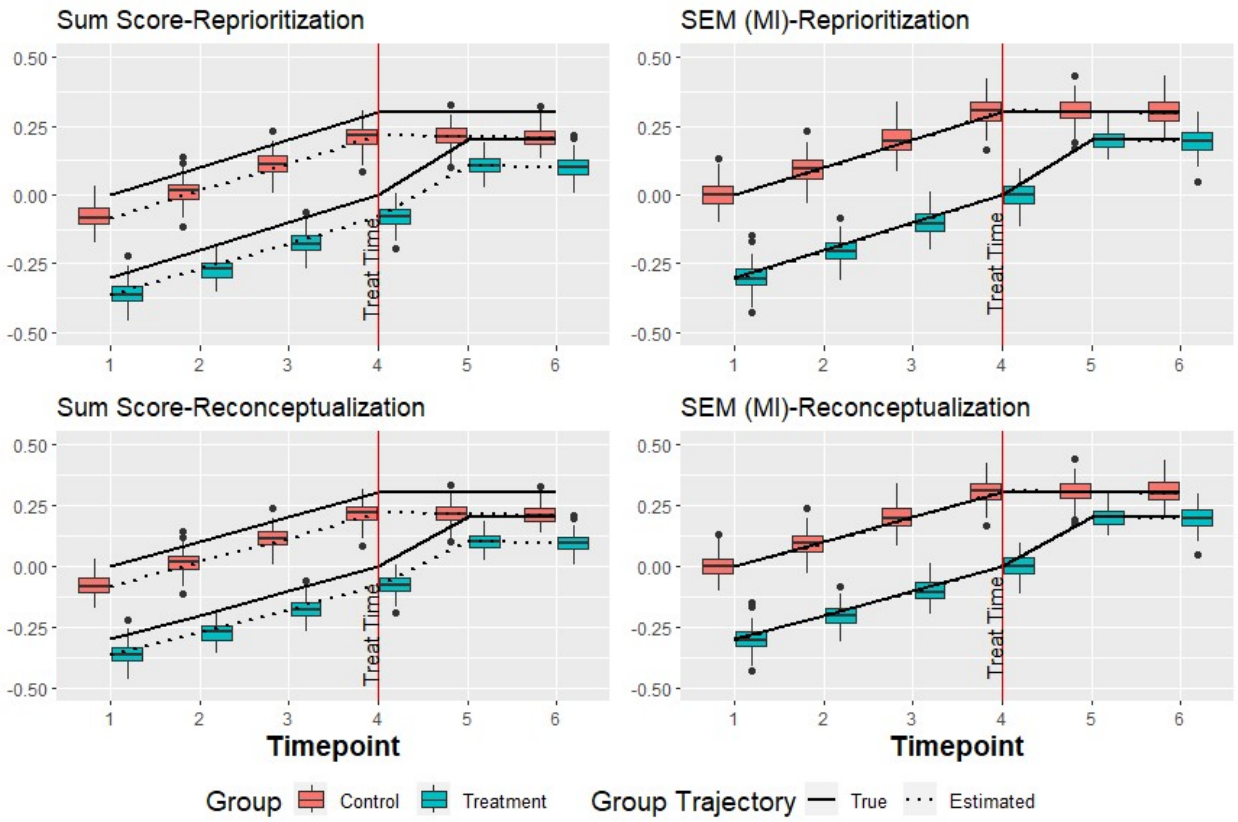

Figure 13. *Boxplots showing means by timepoint across all 100 replications for .2 treatment effect, High Loading, N=500, 8-item condition in Simulation Study 3*

Due to the number of conditions not every result is relayed in this section. A summary of the treatment effect bias for every condition as well as the recovery of loadings are presented in Appendix II and mean timepoints are located online at https://osf.io/jp2xs/.

Discussion

DiD designs can be applied in many contexts in which latent variables are of interest, and generally, these types of variables are measured using multiple items. In particular, survey scales

are a popular way to measure latent constructs, and they are typically aggregated as sum scores, which rely on a number of oftentimes unjustifiable assumptions (McNeish & Wolf, 2020). When using a DiD design these assumptions have a greater potential to be violated due to the differences between control and treatment groups and the measurement across time that happens in DiD. Furthermore, measurement changes have the potential to violate the parallel trend assumption of the DiD design also inducing bias. The results of this study show how these measurement issues can affect a DiD analysis.

First, the simulation results show that, especially when factor loadings are low and there are small number of items in a measure, sum scoring can induce substantial bias into treatment effect estimates. Wrongly fitting a sum score model resulted in understating a true treatment effect of .2 units by anywhere from .01 to .08 units depending on the true loadings in the model. In other words, using a sum score model when its assumptions were violated led to a downward bias of up to 40% of the true treatment effect. Importantly, a larger sample size counterintuitively made the bias worse. This creates tradeoffs for analysts who wish to use sum scores, as a large sample size is desirable for detection of small effects and interactions. In contrast, the bias from SEM was negligible.

There are also implications for the parallel trends assumption because the pre-treatment trend recovery was hampered by the use of sum scoring. The proportional nature of sum score bias means that it can contribute to illusory trends that could mislead an analyst about the viability of the parallel trends assumption. Even though this does not change the validity of the parallel trends assumption itself, it can impact the results of a study through an analyst's behavior. While the differences were small in this simulation, this is likely to be a greater problem when the trends are steeper.

Second, the simulation results show that stable measurement noninvariance between control and treatment groups in a DiD can have substantial impacts on treatment estimates if measurement invariance is assumed. While the form of CT measurement noninvariance induced mostly counteracted the downward bias of sum scoring, this is not guaranteed and requires specific circumstances to work perfectly. Balancing two sources of bias that are in opposite directions means that not accounting for differences in measurement can cause unpredictable problems and it is not safe to assume that sum scoring will always give a conservative estimate of the treatment effect.

For SEM the assumption of measurement invariance inflated the true treatment effect. This led up to a maximum of a 35% overstatement of the true treatment effect with medium CT noninvariance. While this might not be a problem for detection of effects, other types of CT noninvariance are likely to understate the true treatment effect by the same magnitude. Furthermore, the study results show the presence of CT noninvariance has the potential to make the problem of recovery pre-treatment trend recovery worse. Accounting for CT noninvariance could be key to determining if a dataset is usable in a DiD design or if adjustments need to be made to meet the parallel trends assumption.

Finally, when response shifts lead to measurement noninvariance between control and treatment groups in a DiD, ignoring those shifts (as when using a sum score model or a model that assumes invariance) led to the greatest amount of bias in treatment effect estimates. Ignoring the measurement noninvariance led to the true treatment effect being understated by 50% or more in several conditions.

Recommendations for Education Researchers

Given these results there is a strong argument for using scoring strategies that are flexible and have minimal assumptions. This is especially true when the dependent variable is composed from a small number of items and in the presence of measurement noninvariance. Education researchers that are using multi-item outcomes in a DiD design should first carefully consider the nature of their data. If the groups are diverse, or the study time covers developmental periods it is particularly important to avoid sum scoring. Even if the sample size is too small or the analysis too complicated to implement entirely in SEM researchers should consider using factor scoring or similar approaches instead of sum scoring.

If researchers must use sum scores for ease of interpretation or other logistical reasons, researchers should at a minimum determine how well the items load on to the construct. Ideally, they should also test for measurement invariance to understand the possible impact that sum scoring could have on the analysis.

Limitations and Future Directions

This study has a few limitations that are important. First, it is essential to note that this study assumes that variations in measurement arise due to changes in how participants perceive and answer survey questions, and the simulations are based off of this scenario. However, in empirical data it is possible that such differences in measurement are actually signaling other factors that undermine the validity of a DiD design. In those cases, the benefits of SEM outlined in this study are not likely to be realized. Second, while I tried to cover a broad range of simulation conditions, there are many more scenarios to consider. For example, I only examined loading measurement noninvariance. Measurement noninvariance of intercepts or variances could also have important impacts. There are many more factors that could be explored in future research. Furthermore, this study used the most basic of DiD designs and did not include

covariates. The impact of scoring on more complicated DiD designs or those that use covariates need to be explored so that researchers understand the impact of scoring in more complicated DiD designs.

Going forward, the study's results also suggest a greater need for research understanding the tradeoffs between scoring methods when sample sizes are very small. On one hand, the study clearly shows that using a latent variable model can address measurement error and reduce measurement bias, which in turn can bias treatment effect estimates. On the other, sum scores treat many measurement model parameters as fixed, and therefore do not introduce uncertainty into estimates of parameters like loadings. These tradeoffs merit additional research, including in other quasi-experimental contexts and more complicated DiD designs.

Finally, the present study operates on the assumption that the data generating model we use correctly approximates the data generating model for DiD data. Unfortunately, it is impossible to ascertain the exact measurement model of empirical data. As a result, the gains in accuracy observed in our study represent a ceiling. If the true measurement model of empirical data deviates from the data generating model, we used the misspecification may result in bias. Nonetheless, sum scoring is unlikely to match the data generating model better, indicating that the comparative advantages identified in our study will endure.

Chapter 1 References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton university press.

Barnard-Brak, L., Watkins, L., & Richman, D. M. (2021). Autocorrelation and estimates of treatment effect size for single-case experimental design data. *Behavioral Interventions*, *36*(3), 595–605.

Beck, A. T., & Beamesderfer, A. (1974). Assessment of depression: The depression inventory. In *Psychological measurements in psychopharmacology* (Vol. 7, pp. 151–169). Karger Publishers.

Bifulco, R. (2012). Can nonexperimental estimates replicate estimates based on random assignment in evaluations of school choice? A within-study comparison. *Journal of Policy Analysis and Management*, *31*(3), 729–751.

Callaway, B., & Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Cunningham, S. (2020). Causal Inference. *The Mixtape*, *1*.

Dimick, J. B., & Ryan, A. M. (2014). Methods for evaluating changes in health care policy: The difference-in-differences approach. *Jama*, *312*(22), 2401–2402.

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430.

Hallberg, K., Williams, R., & Swanlund, A. (2020). Improving the use of aggregate longitudinal data on school performance to assess program effectiveness: Evidence from three within study comparisons. *Journal of Research on Educational Effectiveness*, *13*(3), 518–545.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in M plus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Research Paper. MET Project. (2010). In *Bill & Melinda Gates Foundation*. Bill & Melinda Gates Foundation. https://eric.ed.gov/?id=ED528382

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.

Muthén, L. K., & Muthén, B. (2017). Mplus user's guide: Statistical analysis with latent variables, user's guide. Muthén & Muthén.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*(3), 587–598.

Parsons, J., & Taylor, L. (2011). Improving student engagement. *Current Issues in Education*, *14*(1).

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2019). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*.

Rocconi, L. M., & Gonyea, R. M. (2018). Contextualizing Effect Sizes in the National Survey of Student Engagement: An Empirical Analysis. *Research & Practice in Assessment*, *13*, 22–38.

Schwartz, C. E., & Sprangers, M. A. (2000). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference/William R. Shadish, Thomas D. Cook, Donald T. Campbell. Boston: Houghton Mifflin.

Sheldon, K. M., & Biddle, B. J. (1998). Standards, accountability, and school reform: Perils and pitfalls. *Teachers College Record*, *100*(1), 164–180.

Snow, J. (1855). On the mode of communication of cholera. John Churchill.

Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). Digest of Education Statistics 2017, NCES 2018-070. *National Center for Education Statistics*.

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, 01466216211013102.

Soland, J. (2022). Evidence That Selecting an Appropriate Item Response Theory–Based Approach to Scoring Surveys Can Help Avoid Biased Treatment Effect Estimates. *Educational and Psychological Measurement*, 00131644211007551.

Soland, J., Johnson, A., & Talbert, E. (2022). Regression Discontinuity Designs in a Latent Variable Framework. *Psychological Methods*.

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation. *MDRC*.

Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: A theoretical model. *Social Science & Medicine*, *48*(11), 1507–1515.

St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, *35*(3), 311–327.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292.

Van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

Wang, M.-T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16–26.

West, M. R., Buckley, K., Krachman, S. B., & Bookman, N. (2018). Development and implementation of student social-emotional surveys in the CORE Districts. *Journal of Applied Developmental Psychology*, *55*, 119–129.

Whitney, C. R., & Candelaria, C. A. (2017). The effects of No Child Left Behind on children's socioemotional outcomes. *AERA Open*, *3*(3), 2332858417726324.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, *39*.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 435–463.

Chapter 1: Appendix I

The effects of treating each timepoint separately can be seen mathematically by modifying the assumptions of equation 5. Removing the assumptions that $\boldsymbol{v}$ is zero and β_0 is zero gives us:

$$y_i = v + \lambda(\beta_0 + \beta_1 a_i + \beta_2 t_i + \beta_3 a_i t_i + \epsilon_i) + \varepsilon_i$$

(1)

Given $E(\epsilon_i) = 0$ and $E(\varepsilon_i) = 0$,

$$E(y_i) = E(v + \lambda(\beta_0 + \beta_1 a_i + \beta_2 t_i + \beta_3 a_i t_i))$$

(2)

When measurement is not invariant then the intercept and loading between the groups are not equivalent and both λ and v require a g subscript. Therefore, for each group the expected value of y_i are as follows:

For the control group before treatment

$$E(y_i \mid a_i = 0 \,\&\, t_i = 0) = v_g + \lambda_g \beta_0$$

(3)

For the control group after treatment

$$E(y_i \mid a_i = 0 \,\&\, t_i = 1) = v_g + \lambda_g \beta_0 + \lambda_g \beta_2 E(t_i) = v_g + \lambda_g \beta_0 + \lambda_g \beta_2$$

(4)

For the treatment group before treatment

$$E(y_i \mid a_i = 1 \,\&\, t_i = 0) = v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1 E(a_i) = \; v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1$$

(5)

For the treatment group after treatment

$$E(y_i \mid a_i = 0 \,\&\, t_i = 1) = \; v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1 E(a_i) + \lambda_g\,\beta_2 E(t_i) + \lambda_g\,\beta_3 E(a_i)E(t_i)$$

$$= \; v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1 + \lambda_g\,\beta_2 + \lambda_g\,\beta_3$$

(6)

Thus the expectation of the observed causal estimate is:

$$E(y_i \mid a_i = 1 \,\&\, t_i = 1) - \; E(y_i \mid a_i = 1 \,\&\, t_i = 0)] - [E(y_i \mid a_i = 0 \,\&\, t_i = 1) -$$

$$E(y_i \mid a_i = 0 \,\&\, t_i = 0) = v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1 + \lambda_g\,\beta_2 + \lambda_g\,\beta_3 - v_g + \lambda_g\beta_0 + \lambda_g\,\beta_1 - v_g +$$

$$\lambda_g\beta_0 + \lambda_g\,\beta_2 - v_g + \lambda_g\beta_0$$

(7)

As seen in Equation 17 the true causal estimate $\beta_3$is not only modified by λ_g as it is in equation 10 but the other coefficients and variations of λ_g and v_g. Measurement invariance means that the observed scores for each group are weighted by their own unique variations of λ and v which can introduce more bias into the causal estimate if not accounted for.

Chapter 1 Appendix II

A.2 Table 1. Simulation Study 1 Bias by Condition

Sample Size	Treatment Effect (SDs)	# Items	Loadings	Factor Score Bias	SEM Bias	Sum Score Bias
500	0	4	High	0.009	-0.001	0.007
500	0	4	Low	0.015	0.002	0.008
500	0	4	Mix	0.011	0.005	0.009
500	0	8	High	0.002	0.008	0.002
500	0	8	Low	0.002	0.008	0.002
500	0	8	Mix	-0.001	0.005	0.001
3700	0	4	High	-0.001	-0.001	0.000
3700	0	4	Low	-0.003	-0.003	-0.001
3700	0	4	Mix	-0.003	-0.001	-0.002
3700	0	8	High	0.000	-0.001	0.000
3700	0	8	Low	-0.001	-0.002	0.000
3700	0	8	Mix	0.002	0.000	0.001
500	0.1	4	High	0.007	-0.001	-0.001
500	0.1	4	Low	0.013	0.002	-0.030
500	0.1	4	Mix	0.009	0.004	-0.024

500	0.1	8	High	0.000	0.008	-0.003
500	0.1	8	Low	0.000	0.008	-0.023
500	0.1	8	Mix	-0.003	0.005	-0.017
3700	0.1	4	High	-0.002	-0.001	-0.009
3700	0.1	4	Low	-0.004	-0.003	-0.039
3700	0.1	4	Mix	-0.005	-0.001	-0.036
3700	0.1	8	High	-0.002	-0.001	-0.005
3700	0.1	8	Low	-0.003	-0.002	-0.024
3700	0.1	8	Mix	0.000	0.000	-0.017
500	0.2	4	High	0.005	-0.001	-0.010
500	0.2	4	Low	0.010	0.002	-0.067
500	0.2	4	Mix	0.006	0.004	-0.058
500	0.2	8	High	-0.002	0.008	-0.009
500	0.2	8	Low	-0.002	0.008	-0.047
500	0.2	8	Mix	-0.005	0.005	-0.035
3700	0.2	4	High	-0.004	0.000	-0.018
3700	0.2	4	Low	-0.006	-0.003	-0.076
3700	0.2	4	Mix	-0.007	-0.001	-0.069
3700	0.2	8	High	-0.004	-0.001	-0.010

3700	0.2	8	Low	-0.005	-0.002	-0.048
3700	0.2	8	Mix	-0.002	0.000	-0.036

A.2 Table 2. Simulation Study 2 Bias by Condition (Small CT Noninvariance)

Sample Size	Treatment Effect (SDs)	# Items	Loadings	Factor Score Bias	SEM (CT) Bias	SEM (MI) Bias	Sum Score Bias
500	0	4	High	0.009	-0.001	-0.001	0.016
500	0	4	Low	0.014	0.003	0.000	0.021
500	0	4	Mix	0.012	0.005	0.004	0.022
500	0	8	High	0.003	0.008	0.008	0.007
500	0	8	Low	0.003	0.009	0.009	0.009
500	0	8	Mix	0.001	0.007	0.007	0.008
3700	0	4	High	0.000	0.000	0.000	0.009
3700	0	4	Low	-0.001	-0.002	-0.003	0.011
3700	0	4	Mix	-0.003	-0.001	-0.001	0.010
3700	0	8	High	0.001	-0.001	-0.001	0.005
3700	0	8	Low	0.000	-0.002	-0.002	0.008

3700	0	8	Mix	0.004	0.000	0.000	0.008
500	0.1	4	High	0.006	-0.001	0.001	0.011
500	0.1	4	Low	0.011	0.002	0.005	-0.010
500	0.1	4	Mix	0.009	0.005	0.009	-0.005
500	0.1	8	High	0.001	0.009	0.010	0.004
500	0.1	8	Low	0.001	0.010	0.012	-0.011
500	0.1	8	Mix	-0.001	0.007	0.009	-0.007
3700	0.1	4	High	-0.003	0.000	0.002	0.004
3700	0.1	4	Low	-0.004	-0.002	0.002	-0.019
3700	0.1	4	Mix	-0.005	-0.001	0.004	-0.017
3700	0.1	8	High	-0.001	-0.001	0.000	0.002
3700	0.1	8	Low	-0.002	-0.002	0.000	-0.013
3700	0.1	8	Mix	0.002	0.000	0.002	-0.007
500	0.2	4	High	0.002	-0.001	0.003	0.006
500	0.2	4	Low	0.006	0.002	0.010	-0.041
500	0.2	4	Mix	0.005	0.005	0.014	-0.033
500	0.2	8	High	-0.003	0.009	0.011	0.000
500	0.2	8	Low	-0.003	0.010	0.015	-0.032
500	0.2	8	Mix	-0.005	0.007	0.011	-0.022

3700	0.2	4	High	-0.007	0.000	0.004	-0.001
3700	0.2	4	Low	-0.008	-0.002	0.007	-0.050
3700	0.2	4	Mix	-0.009	-0.001	0.009	-0.044
3700	0.2	8	High	-0.005	-0.001	0.001	-0.002
3700	0.2	8	Low	-0.006	-0.002	0.003	-0.034
3700	0.2	8	Mix	-0.002	0.000	0.005	-0.023

A.2 Table 3. Simulation Study 2 Bias by Condition (Medium CT Noninvariance)

Sample Size	Treatment Effect (SDs)	# Items	Loadings	Factor Score Bias	SEM (CT) Bias	SEM (MI) Bias	Sum Score Bias
500	0	4	High	0.008	-0.001	-0.001	0.028
500	0	4	Low	0.014	0.002	-0.002	0.038
500	0	4	Mix	0.011	0.006	0.004	0.039
500	0	8	High	0.005	0.008	0.009	0.015
500	0	8	Low	0.004	0.011	0.011	0.021
500	0	8	Mix	0.002	0.009	0.009	0.019
3700	0	4	High	0.001	0.001	0.001	0.022

3700	0	4	Low	0.000	-0.002	-0.002	0.029
3700	0	4	Mix	-0.003	-0.001	-0.001	0.027
3700	0	8	High	0.002	0.000	0.000	0.013
3700	0	8	Low	0.001	-0.002	-0.001	0.018
3700	0	8	Mix	0.004	0.000	0.000	0.018
500	0.1	4	High	0.005	0.000	0.005	0.029
500	0.1	4	Low	0.011	0.002	0.011	0.017
500	0.1	4	Mix	0.008	0.006	0.015	0.021
500	0.1	8	High	0.002	0.009	0.013	0.015
500	0.1	8	Low	0.002	0.011	0.018	0.005
500	0.1	8	Mix	-0.001	0.009	0.016	0.008
3700	0.1	4	High	-0.002	0.001	0.007	0.022
3700	0.1	4	Low	-0.003	-0.002	0.010	0.008
3700	0.1	4	Mix	-0.006	-0.001	0.010	0.010
3700	0.1	8	High	0.000	0.000	0.004	0.013
3700	0.1	8	Low	-0.001	-0.001	0.006	0.003
3700	0.1	8	Mix	0.002	0.000	0.007	0.007
500	0.2	4	High	0.001	0.000	0.011	0.029
500	0.2	4	Low	0.007	0.002	0.023	-0.004

500	0.2	4	Mix	0.004	0.007	0.026	0.003
500	0.2	8	High	-0.001	0.010	0.017	0.015
500	0.2	8	Low	-0.002	0.011	0.026	-0.011
500	0.2	8	Mix	-0.004	0.009	0.023	-0.003
3700	0.2	4	High	-0.006	0.001	0.014	0.022
3700	0.2	4	Low	-0.006	-0.002	0.023	-0.014
3700	0.2	4	Mix	-0.009	-0.001	0.021	-0.008
3700	0.2	8	High	-0.003	0.001	0.008	0.013
3700	0.2	8	Low	-0.005	-0.001	0.013	-0.013
3700	0.2	8	Mix	-0.001	0.000	0.013	-0.004

A.2 Table 4. Simulation Study 3 Bias by Condition (Reprioritization)

Sample Size	Treatment Effect (SDs)	# Items	Loadings	Factor Score Bias	SEM (MI) Bias	SEM (RS) Bias	Sum Score Bias
500	0	4	High	0.008	0.000	-0.001	0.007
500	0	4	Low	0.019	0.003	0.002	0.008
500	0	4	Mix	0.014	0.006	0.008	0.009

500	0	8	High	0.001	0.007	0.007	0.002
500	0	8	Low	0.001	0.006	0.005	0.001
500	0	8	Mix	-0.003	0.003	0.003	0.000
3700	0	4	High	-0.002	-0.001	-0.001	0.000
3700	0	4	Low	-0.002	-0.003	-0.002	-0.001
3700	0	4	Mix	-0.015	-0.001	-0.003	-0.002
3700	0	8	High	-0.001	-0.001	-0.002	0.000
3700	0	8	Low	-0.001	-0.003	-0.003	0.000
3700	0	8	Mix	0.001	0.000	0.000	0.001
500	0.1	4	High	0.007	-0.004	-0.001	-0.012
500	0.1	4	Low	0.015	-0.009	0.002	-0.046
500	0.1	4	Mix	-0.006	-0.027	-0.001	-0.040
500	0.1	8	High	0.000	0.006	0.008	-0.009
500	0.1	8	Low	0.000	0.001	0.005	-0.032
500	0.1	8	Mix	-0.004	-0.005	0.003	-0.026
3700	0.1	4	High	-0.003	-0.004	0.000	-0.020
3700	0.1	4	Low	-0.004	-0.015	-0.002	-0.055
3700	0.1	4	Mix	-0.022	-0.034	-0.002	-0.052
3700	0.1	8	High	-0.002	-0.003	-0.002	-0.010

3700	0.1	8	Low	-0.002	-0.008	-0.003	-0.033
3700	0.1	8	Mix	-0.001	-0.008	0.000	-0.026
500	0.2	4	High	0.004	-0.008	-0.001	-0.032
500	0.2	4	Low	0.010	-0.021	0.003	-0.099
500	0.2	4	Mix	-0.026	-0.058	-0.011	-0.090
500	0.2	8	High	-0.002	0.005	0.008	-0.019
500	0.2	8	Low	-0.001	-0.004	0.006	-0.065
500	0.2	8	Mix	-0.007	-0.013	0.003	-0.053
3700	0.2	4	High	-0.005	-0.008	0.000	-0.039
3700	0.2	4	Low	-0.006	-0.027	-0.002	-0.108
3700	0.2	4	Mix	-0.029	-0.066	-0.002	-0.101
3700	0.2	8	High	-0.004	-0.004	-0.002	-0.020
3700	0.2	8	Low	-0.004	-0.014	-0.003	-0.066
3700	0.2	8	Mix	-0.003	-0.016	0.000	-0.053

A.2 Table 5. Simulation Study 3 Bias by Condition (Reconceptualization)

Sample Size	Treatment Effect (SDs)	# Items	Loadings	Factor Score Bias	SEM (MI) Bias	SEM (RS) Bias	Sum Score Bias
500	0	4	High	0.009	0.000	-0.001	0.007

500	0	4	Low	0.017	0.003	0.002	0.008
500	0	4	Mix	0.023	0.007	0.001	0.009
500	0	8	High	0.001	0.007	0.007	0.001
500	0	8	Low	0.001	0.006	0.006	0.001
500	0	8	Mix	-0.003	0.002	0.000	0.000
3700	0	4	High	-0.002	0.000	0.000	0.000
3700	0	4	Low	-0.004	-0.003	-0.002	-0.001
3700	0	4	Mix	-0.005	-0.001	-0.001	-0.002
3700	0	8	High	-0.001	-0.001	-0.002	0.000
3700	0	8	Low	-0.002	-0.003	-0.003	0.000
3700	0	8	Mix	0.001	0.000	0.000	0.001
500	0.1	4	High	0.008	-0.005	-0.001	-0.020
500	0.1	4	Low	0.012	-0.007	0.001	-0.042
500	0.1	4	Mix	-0.002	-0.049	-0.022	-0.053
500	0.1	8	High	0.000	0.005	0.007	-0.013
500	0.1	8	Low	0.000	0.002	0.006	-0.030
500	0.1	8	Mix	-0.004	-0.010	0.001	-0.033
3700	0.1	4	High	-0.002	-0.005	0.000	-0.028
3700	0.1	4	Low	-0.006	-0.013	-0.002	-0.051

3700	0.1	4	Mix	-0.013	-0.058	-0.002	-0.064
3700	0.1	8	High	-0.002	-0.004	-0.002	-0.014
3700	0.1	8	Low	-0.003	-0.007	-0.003	-0.031
3700	0.1	8	Mix	0.000	-0.012	-0.001	-0.032
500	0.2	4	High	0.005	-0.010	-0.001	-0.048
500	0.2	4	Low	0.008	-0.016	0.002	-0.091
500	0.2	4	Mix	-0.026	-0.104	-0.034	-0.115
500	0.2	8	High	-0.002	0.003	0.008	-0.028
500	0.2	8	Low	-0.002	-0.002	0.007	-0.061
500	0.2	8	Mix	-0.007	-0.022	0.001	-0.066
3700	0.2	4	High	-0.004	-0.011	0.000	-0.055
3700	0.2	4	Low	-0.009	-0.022	-0.002	-0.100
3700	0.2	4	Mix	-0.021	-0.113	-0.002	-0.126
3700	0.2	8	High	-0.004	-0.006	-0.002	-0.028
3700	0.2	8	Low	-0.005	-0.011	-0.003	-0.062
3700	0.2	8	Mix	-0.003	-0.024	-0.001	-0.065

A.2 Table 6. Loading Recovery Check

Sample Size	# Items	Loadings	Measurement Invariance	Recovered Loading Bias
3700	4	High	Invariant	-0.001
500	4	High	Invariant	0.000
3700	4	Low	Invariant	-0.001
500	4	Low	Invariant	0.001
3700	4	Mix	Invariant	-0.001
500	4	Mix	Invariant	0.000
3700	8	High	Invariant	-0.001
500	8	High	Invariant	-0.001
3700	8	Low	Invariant	0.000
500	8	Low	Invariant	0.000
3700	8	Mix	Invariant	0.000
500	8	Mix	Invariant	0.000
3700	4	High	Medium CT	-0.001
500	4	High	Medium CT	-0.001
3700	4	Low	Medium CT	0.000
500	4	Low	Medium CT	0.001
3700	4	Mix	Medium CT	0.000

500	4	Mix	Medium CT	0.002
3700	8	High	Medium CT	-0.001
500	8	High	Medium CT	0.000
3700	8	Low	Medium CT	-0.001
500	8	Low	Medium CT	-0.001
3700	8	Mix	Medium CT	0.000
500	8	Mix	Medium CT	-0.002
3700	4	High	Reconceptualization	-0.001
500	4	High	Reconceptualization	0.000
3700	4	Low	Reconceptualization	-0.004
500	4	Low	Reconceptualization	-0.003
3700	4	Mix	Reconceptualization	0.004
500	4	Mix	Reconceptualization	1.912
3700	8	High	Reconceptualization	-0.001
500	8	High	Reconceptualization	-0.001
3700	8	Low	Reconceptualization	-0.001
500	8	Low	Reconceptualization	0.000
3700	8	Mix	Reconceptualization	-0.001
500	8	Mix	Reconceptualization	0.000

3700	4	High	Reprioritization	-0.001
500	4	High	Reprioritization	0.000
3700	4	Low	Reprioritization	-0.002
500	4	Low	Reprioritization	-0.004
3700	4	Mix	Reprioritization	-0.004
500	4	Mix	Reprioritization	-0.079
3700	8	High	Reprioritization	-0.001
500	8	High	Reprioritization	-0.001
3700	8	Low	Reprioritization	-0.001
500	8	Low	Reprioritization	0.000
3700	8	Mix	Reprioritization	0.000
500	8	Mix	Reprioritization	0.000
3700	4	High	Small CT	0.000
500	4	High	Small CT	0.000
3700	4	Low	Small CT	0.000
500	4	Low	Small CT	0.001
3700	4	Mix	Small CT	0.000
500	4	Mix	Small CT	0.002
3700	8	High	Small CT	-0.001

500	8	High	Small CT	-0.001
3700	8	Low	Small CT	-0.001
500	8	Low	Small CT	-0.001
3700	8	Mix	Small CT	0.000
500	8	Mix	Small CT	-0.003

Chapter 2: Can Scoring Decisions Affect Results from Quasi-experimental Studies? Revisiting Effects of No Child Left Behind on Children's Socioemotional Outcomes

Chapter 2 Abstract

The Difference-in-Difference (DiD) design is a useful tool for evaluating the impact of policies on various outcomes. It accounts for differences between treated and untreated groups before treatment by comparing their outcomes over time. Educational evaluators use DiD designs to produce unbiased and precise estimates, including outcomes related to students' academic, psychological, and socio-emotional attributes measured using surveys and tests. The use of survey-based outcomes in DiD designs is growing due to the increasing availability of socio-emotional learning (SEL) data and popularity of SEL policies. However, a challenge in using survey-based outcomes is scoring the item responses, which can affect the causal estimates. Most related work on scoring has been conducted using simulated data and not in the context of DiD designs. This research aims to investigate the impact of scoring decisions in a DiD context by rescoring a study by Whitney and Candelaria (2017).

Keywords: measurement; item response theory; difference-in-difference designs; program evaluation.

Can Scoring Decisions Affect Results from Quasi-experimental Studies?

Revisiting Effects of No Child Left Behind on Children's Socioemotional Outcomes

The Difference-in-Difference (DiD) design is one of the most useful designs to evaluate the causal impact of education policies. The DiD design works by accounting for the differences between treated and untreated groups pre-intervention when comparing the outcomes of those groups. The core logic of DiD is that any differences between the treatment and non-treatment groups will remain consistent over time, allowing for an unbiased causal estimate (Caniglia & Murray, 2020). Using its relatively simple and flexible design, educational evaluators can, under the right assumptions, produce unbiased and precise causal estimates (Somers et al., 2013). The most basic DiD design only needs data from before and after a policy for two groups, a group affected by a policy and a group that has not been affected (Wing et al., 2018). This basic design can then be modified to account for more complex situations, such as multiple groups with differential treatment timings (Callaway & Sant'Anna, 2020).

DiD designs are often used to estimate the impact of polices on easily quantifiable outcomes like funding levels (Delaney & Kearney, 2015). However, in education, there is also interest in outcomes related to students' academic, psychological and socio-emotional attributes, which are latent and often measured using surveys and tests. For instance, outcomes like mathematics engagement are of interest to educational researchers and are regularly studied (e.g. Fredricks et al., 2018; Talbert et al., 2019). While few studies have used these outcomes in a DiD design, related studies are starting to emerge given the growing interest in socio-emotional learning (SEL) outcomes in education (Gehlbach & Hough, 2018). For example, under the 2017 *Every Student Succeeds Act*, states are able to use funds under Title I and Title II, Part A to address SEL (Grant et al., 2017). At the state level (Eklund et al., 2018) and local levels (Mahoney et al.,

2020) policies and programs are being created and implemented to improve socio-emotional outcomes and more data are being collected about them. For example, CORE, a collaboration of several districts with over a million children in California, uses SEL outcomes as improvement measures and collects SEL data annually (Gehlbach & Hough, 2018).

The increasing availability of SEL data and popularity of SEL policies drives the opportunity and need for using survey-based outcomes in DiD designs. A recent example of such a study is one conducted by Whitney and Candelaria (2017) that uses a DiD design to investigate the impact of The No Child Left Behind (NCLB) Act of 2001 on various socio-emotional outcomes such as mathematics interest. In that study, they found that the NCLB had both positive and negative effects on socio-emotional outcomes, with possible improvements in math competence and interest, but also an increase in academic anxiety. Their subgroup analysis suggested that the treatment effect on math interest and competence varied by SES and sex, with significant positive effects for students from lower SES backgrounds, male students for math interest, and female students for math competence. However, a challenge with the use of survey-based outcomes in evaluation is how to score the survey item responses. In many studies, including Whitney and Candelaria (2017), scores are often produced by taking an average or the sum of item responses, known as sum scoring (McNeish & Wolf, 2020). The frequent use of sum scoring occurs despite research demonstrating that scoring approaches (e.g., the model used to calibrate item parameters and score the item responses, as well as the approach to producing those scores in an item response theory [IRT] context) can bias the treatment estimates under realistic conditions in a variety of designs. For example, studies examining randomized control trials (Gorter et al., 2016; Soland, 2021, 2022), regression discontinuity designs (Soland, Johnson, et al., 2022), and longitudinal designs more generally (Bauer & Curran, 2015; Gorter et

al., 2015; Kuhfeld & Soland, 2020; Soland, Kuhfeld, et al., 2022) all used conditions that were based on the results of empirical studies or relevant literature. This body of work clearly suggests that scoring decisions—including measurement model misspecification—could impact causal estimates derived from the DiD design.

However, most related work on how scoring affects inferences drawn in program evaluation contexts has two limitations in terms of understanding scoring in a DiD context. First, almost all of the analyses were conducted using simulated data, not empirical data[1]. Thus, the practical implications are somewhat murky. Second, the effect of scoring has not been investigated in the DiD context specifically, only in the context of other quasi-experimental methods like regression discontinuity (Soland, Johnson, et al., 2022). To make the impact of scoring decisions more concrete in this context, work specific to the DiD design using real data is needed.

To that end, this study investigates how different IRT scoring approaches affect the substantive findings from Whitney and Candelaria (2017), and specifically how these approaches affect our understanding of how NCLB impacted students' socio-emotional outcomes. We decided to re-examine the evaluation by Whitney and Candelaria (2017) because it was a comprehensive DiD evaluation of an important policy that analyzed several outcomes and discovered both non-significant and significant statistical results. Further, these outcomes were produced by using rudimentary scoring approaches that can lead to biased treatment effect estimates (e.g., Gorter et al., 2020). An additional benefit of revisiting Whitney and Candelaria is that it uses a well-known educational dataset, the Early Childhood Longitudinal Study Kindergarten-Fifth Grade (ECLS-K) (Tourangeau et al., 2009), which means other researchers can use our code to score measures

[1] While some studies do include empirical demonstrations these are typically only brief, illustrative empirical examples.

of interest from the dataset, to understand how common uses of those scores are impacted by measurement decisions. By re-examining the study, we are able to investigate the following research questions:

1. Do conclusions about the effects of accountability via NCLB on children's socio-emotional outcomes change dependent on which scoring approach is used?
2. Are conclusions about different effects by socioeconomic status and sex sensitive to the scoring approach is used?

Background

Existing Knowledge on How Measurement Decisions Can Affect Study Results

There are several decisions that go into scoring a measure, especially in an evaluation context. These decision stages were enumerated previously by Soland, Kuhfeld, and Edwards (2022). By virtue of the effects such decisions can have on bias and variability in observed scores, those decisions can in turn have impacts on the analyses conducted using them. We briefly detail those decision stages here.

Step 1. Deciding whether to use a measurement model versus sum scores. The first decision revolves around choosing whether to use an explicit measurement model. Sum scoring, which includes the common strategy of taking the sum or average of the item responses, does not have an explicit measurement model. However, researchers have argued that it involves an implicit measurement model assuming that every item is equally representative of the measured construct (i.e., that all discrimination/slope parameters are identical) and that those parameters are invariant over time (McNeish & Wolf, 2020). In contrast, more complex statistical approaches, like those based in IRT, estimate a measurement model that allows for items to have unique discrimination/slope parameters, including over time for the same item. Thus, one could

argue that, while sum scores do not involve an overt measurement model, they are equivalent to fitting an extremely constrained measurement model that imposes large assumptions relative to, say, the 2 parameter logistic (2PL) IRT model. Further, when such assumptions are violated, common uses of longitudinal scores relevant to DiD can lead to biased estimates. For instance, Kuhfeld and Soland (2020), showed that, for growth modeling under certain scenarios, using sum scores led to an understatement of true latent slope parameter means and variances by nearly 50%. Similarly, for evaluation studies, treatment effect estimates can be understated by up to 40% in a regression discontinuity design (Soland, Johnson, et al., 2022) and up to 25% in an experiment (Soland, 2021) when using sum scores.

Step 2. Deciding on the specific measurement model. After deciding an explicit measurement model is appropriate, there are a number of choices that can be made around the specific measurement model used. One of the most important is the decision about how to calibrate the item parameters. There are several possible measurement models and, in the case of longitudinal data, samples that could be used to calibrate item parameters in a measurement model. For example, if a researcher was conducting a longitudinal study with children in middle school, they could choose to calibrate using only a cross-section of students from different grades (akin to vertical scaling), based on the first timepoint only, or using a multi-timepoint multidimensional IRT (MIRT) model to calibrate all the item responses at once. More complex calibration requires more data, but choosing a calibration method that does not align with the data generating process can lead to understating parameters by around 20% in growth modeling (Kuhfeld & Soland, 2020) and over 50% in some experimental scenarios (Soland et al., 2022). This bias occurs because of a mismatch between the measurement model and the data-generating process. For example, using a unidimensional IRT model with longitudinal data includes no

explicit correlation of scores over time in the model, and assumes there is only a single mean and variance in the population as opposed to time-specific means and variances. Neither assumption is likely to be met, which potentially leads to bias.

Step 3. Producing scores. Finally, after the model is selected, a researcher must decide how to estimate scores from the measurement model. Researchers can typically use a frequentist approach like MLE (Maximum Likelihood Estimation) or Bayesian approaches like EAP (Expected A Posteriori) or MAP (Maximum A Posteriori). Choosing between MLE versus a Bayesian approach like EAP can involve complex tradeoffs. For instance, MLE scores are asymptotically unbiased and their standard errors are associated with the information function (Baker, 1992). However, MLE cannot easily produce scores for measures where only a single response category is used (Soland, Kuhfeld, et al., 2022). In these cases, the scores are undefined and must be replaced by arbitrary maximums or minimums or left missing. This issue is especially problematic when short self-report measures (like those used to capture SEL outcomes) are employed because respondents frequently use only the top response category of the Likert scale (Soland & Kuhfeld, 2020).

Bayesian approaches also have their own tradeoffs. Bayesian methods incorporate information about the population through the specification of a prior distribution of scores, either using something generic like the standard normal distribution or factoring in something we might know about the construct prior to the study (Bock & Mislevy, 1982). In EAP, the standard normal distribution is typically used, and the estimates of the latent construct are shrunk toward the population mean. Shrinking towards this mean can improve accuracy if the mean aligns with the parameter the researcher is trying to estimate (Soland, Kuhfeld, et al., 2022), but can induce bias if the measurement model does not align with the study design. For example, applying a

single group IRT model using EAP to a scenario with two groups—such as control and treatment conditions—will result in scores being shrunk to a common mean. That is, the model assumes that control and treatment groups are exchangeable. In so doing, the estimated difference between the two groups can be biased downwards because EAP would shrink the scores of the two groups to an overall mean somewhere in between their group-specific means.

Measurement Decisions in Difference in Difference Studies

These measurement decisions could affect the DiD design in many of the same ways they do other study designs. In particular, decisions made in Step 2 (Calibration) are key to the DiD design. The DiD design hinges on an assumption about both changes over time (pre/post intervention) and group differences in changes over time. Specifically, the key DiD assumption, the parallel trends assumption, posits that important unmeasured variables are either time-invariant group attributes or time-varying factors that are group invariant (Wing et al., 2018). When the parallel trends assumption holds, theoretically, accounting factors that might affect the causal estimate are differenced out and the basic DiD design produces an unbiased causal estimate.

However, if the measurement model is not calibrated on the control and treatment groups by timepoint, any differences in the means between groups or over time are not incorporated into the measurement model. If a measurement model disregards either time or group, it implies that the control and treatment groups, or the timepoints, can be interchanged. These are hidden assumptions different from the DiD identification assumptions that are more well-known but are also important. For example, even if the parallel trends assumption, the assumption that important unmeasured variables are either time-invariant group attributes or time-varying factors that are group invariant (Wing et al., 2018), is met perfectly it does not imply that the groups are

exchangeable because even though the group and time attributes are accounted for in the estimating model they are not accounted for in the measurement model.

Unfortunately, both the exchangeability of groups and timepoints are likely untenable assumptions for a DiD design because the design features distinct groups that are observed over time. Unlike in an experiment there is no expectation that the treated and untreated groups are statistically equivalent and the design is often used for situations where there are known differences between the groups on the target construct. Similarly, DiD designs, especially those in education, often take place over long periods of time. Developmental changes or random events like a social media craze may affect the target construct making an assumption of exchangeability between timepoints unsustainable. While for the parallel trends assumption these events need to have a *differential* effect by time or group to constitute a violation, any difference between the groups or timepoints will affect an assumption of exchangeability.

Findings from Whitney and Candelaria

In their study, Whitney and Candelaria used subscales from the children's Self-Descriptive Questionnaire (SDQ), an instrument supported by validity evidence for assessing children's self-concept according to field testing (Atkins-Burnett & Meisels, 2001; Pollack et al., 2005), to create measures of ten socio-emotional outcomes. There are several reasons the use of sum scoring in this context could be problematic. First, each measure was short, composed from only two to five items. Shorter scales can result in more bias because the weighting of each individual item has more importance magnifying the impact of incorrect weightings (Kuhfeld & Soland, 2020; Soland, 2021, 2022; Soland, Kuhfeld, et al., 2022). Second, the measures all involved self-report. Self-report items are vulnerable to differences in interpretation by respondents that could affect measurement. Third, short self-report measures make decisions about which measurement

model to use, as well as whether to use MLE versus EAP scoring, especially consequential (Soland, Kuhfeld, & Edwards, 2022). For example, shrinkage will be greater for shorter, less reliable measures; therefore, decisions about whether to use a unidimensional, versus multi-timepoint multigroup MIRT model that better matches the data can be more impactful.

Measurement issues aside, Whitney and Candelaria (2017) found both beneficial and detrimental effects of NCLB on socio-emotional outcomes[2]. Overall, their analysis suggested that NCLB caused an increase in academic anxiety and possible improvements in math competence and interest. These impacts were relatively small. The increase in academic anxiety was estimated to range from .08 to .14 standard deviations (SDs) while the increase in math competence and interest were estimated to range from .06 to .07 and .05 to .06 SDs respectively. In every model specification, neither the increase in academic anxiety nor the increases in math was statistically significant at the .1 level, but the point estimates were consistent. For the other seven outcomes, none of the model specifications produced estimates that were significant at the .1 level.

Subgroup Findings. Whitney and Candelaria (2017) also conducted a subgroup analysis using the SES and sex variables. These analyses revealed differences in subgroups in the self-reported constructs of math competence and interest by SES and sex. NCLB was estimated to increase math interest by .093 SDs and math competence by .086 SDs in the bottom half of the SES distribution. In comparison, there was an estimated treatment effect of -.001 SDs (math interest) and .027 SDs (math competence) in the top half of the SES distribution. The treatment

[2] Results from Whitney and Candelaria (2017) are exact matches for the sum score results in Appendix I & II

effect was statistically significant at the .05 alpha level for the bottom half of SES distribution and not significant for the top half.

There were also less dramatic differences in math interest and competence by sex. The treatment effect of NCLB on math interest appeared to be greater for male students (.067 SDs, $p<.1$) than female students (.023 SDs). The reverse was true for math competence with a treatment effect of .042 SDs for male students and .071 SDs for female students ($p<.1$)[3].

Methods

Sample

Like the original Whitney and Candelaria (2017) study, this study uses the restricted version of the ECKLS-K dataset, which is nationally representative of kindergartners in 1998–1999 school year. Data were collected in the fall and spring of the kindergarten year, spring of the first-grade year (with a fall subsample), and spring of the third-, fifth-, and eighth-grade years. Like the original study, ours only used data from spring of the first and third grade years, which occurred immediately before and after the passage of NCLB.

IRT Approaches for Item Parameter Calibration

After deciding whether to use a sum score (Step 1), and assuming one prefers an IRT model of some kind, the next decisions for scoring a multi-group, multi-timepoint survey like the one used in Whitney and Candelaria's (2017) DiD study, are the calibration and scoring approaches (Steps 2 and 3). We used several such approaches, including a simplistic calibration approach employing a unidimensional model, as well as an approach more likely to match the data. More

[3] For both the subgroup and total population analysis the authors were concerned that the multiple comparisons made meant that the evidence that they produced was not conclusive.

precisely, the first calibration approach used a unidimensional model to calibrate item parameters using both groups and both timepoints, while in the second, all item parameters were calibrated using a longitudinal multigroup MIRT model that allowed population means and variances to differ by timepoint and group.

Given the original study used Likert-type items, for every scoring approach in our own study, item calibration was accomplished using some version of the graded response model or GRM (Samejima, 1969). We outline this model in the following equations. Let there be $j = 1, \dots, n$ items and $i = 1, \dots, N$ individuals. Let the response from individual i to item j at timepoint t be y_{tij}, where y_{tij} has K response categories. It can be assumed that y_{tij} takes integer values from $(0, \dots, K-1)$. Let the cumulative category response probabilities be

$$P(y_{tij} \geq 1 | \theta_i) = \frac{1}{1 + \exp[-(c_{j1} + a_j \theta_i)]}$$

$$\vdots \qquad (1)$$

$$P(y_{tij} \geq K - 1 | \theta_i) = \frac{1}{1 + \exp[-(c_{j,K-1} + a_j \theta_i)]}$$

The category response probability is the difference between two adjacent cumulative probabilities

$$P(y_{tij} = k | \theta_i) = P(y_{tij} \geq k | \theta_i) - P(y_{tij} \geq k + 1 | \theta_i), \qquad (2)$$

where $\mathrm{P}(y_{tij} \geq 0 | \theta_i)$ is equal to 1 and $\mathrm{P}(y_{tij} \geq K | \theta_i)$ is zero. The item parameter a_j is the slope parameter describing the relationship between item j and the latent factor and $b_{\mathrm{j}1}, \dots, b_{j,K-1}$ are a

set of $K-1$ (strictly ordered) parameters. The thresholds denote the point on the latent variable separating category k from category $k+1$.

In the unidimensional case, the logit in Equation 1 can be re-expressed in a more convenient slope-threshold form as $c_{jk} + a_j\theta_i = a_j(\theta_i - b_{jk})$, where $b_{jk} = -c_{jk}/a_j$ is the threshold (also referred to as severity or difficulty) parameter for category k. The kth threshold denotes the point on the latent variable separating category k from category $k+1$. However, the slope-threshold form does not generalize well to multidimensional models, so we adopted the slope–intercept parameterization for every IRT model that we used. Next, we discuss the specific versions of the GRM used to calibrate item parameters for the survey item responses.

Approach 1. Groups Pooled, Pooled Timepoints, Unidimensional IRT Model. The first approach we used to calibrate the item data was by pooling groups and timepoints. The model was fit with both control and treatment participants at both timepoints, scored all together and ignoring the dependence among item responses from the same person a shown in Figure 1. The item parameters were estimated using a unidimensional IRT model and then the parameters were used to score all item responses in both timepoints and groups. Again, as described in the background section, such an approach has many limitations, including assuming that treatment and control groups are exchangeable in the population (Gorter et al., 2015; Kuhfeld & Soland, 2020).

Approach 2. Groups Unpooled, Separate Timepoints, Multigroup MIRT Model. In a DiD design, researchers are not only interested in changes in time (pre/post intervention); they are interested in group differences in changes over time. Thus, we fit a multigroup, multi-timepoint MIRT model shown in Figure 2. This model allows one to relax assumptions including equality of the latent means and variances across groups/over time (exchangeability). This calibration

approach provides the most flexibility and, perhaps, best matches the nature of DiD data. For example, such a model would allow pre/post scores to have different covariances between control and treatment groups, as well as different variances at one or both timepoints. This model would also allow for different latent means pre/post treatment for both groups.

MIRT models can be further expanded by incorporating demographic subgroups into the model as separate groups for each timepoint as shown in Figure 3. Doing this helps model potential treatment effect heterogeneity for those demographic subgroups by allowing each subgroup at each timepoint to have a different latent mean. If a treatment results in a mean change for a particular demographic subgroup but not another, treating subgroups separately in the measurement model prevents shrinkage towards the total population mean at a given timepoint.

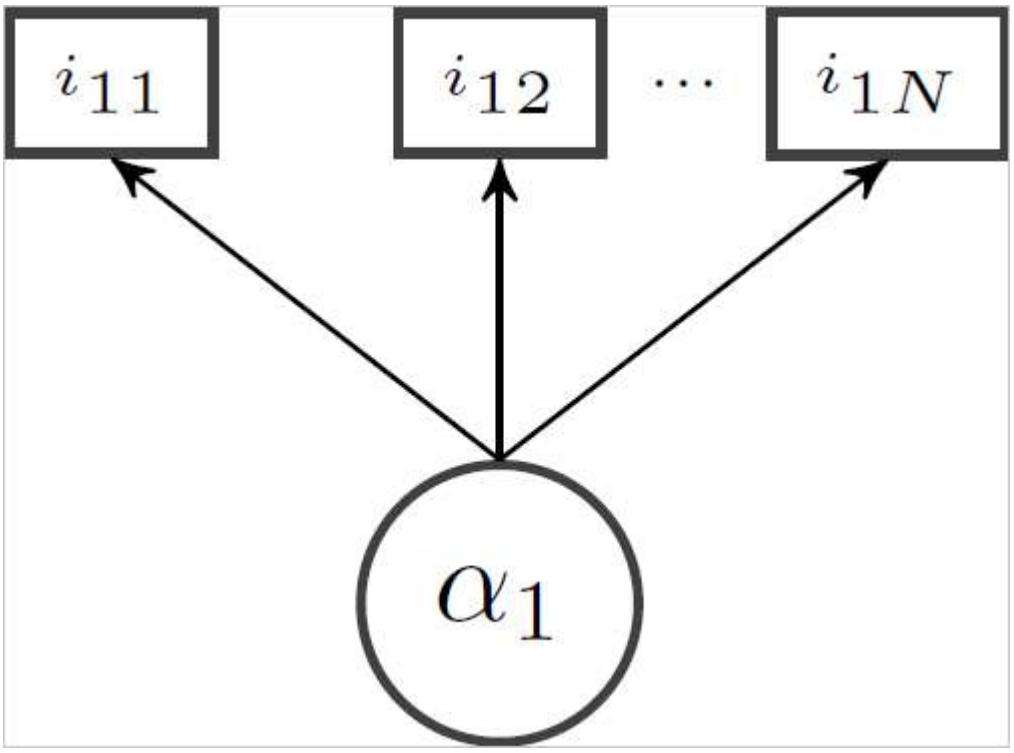

Figure 1. *Unidimensional IRT model*

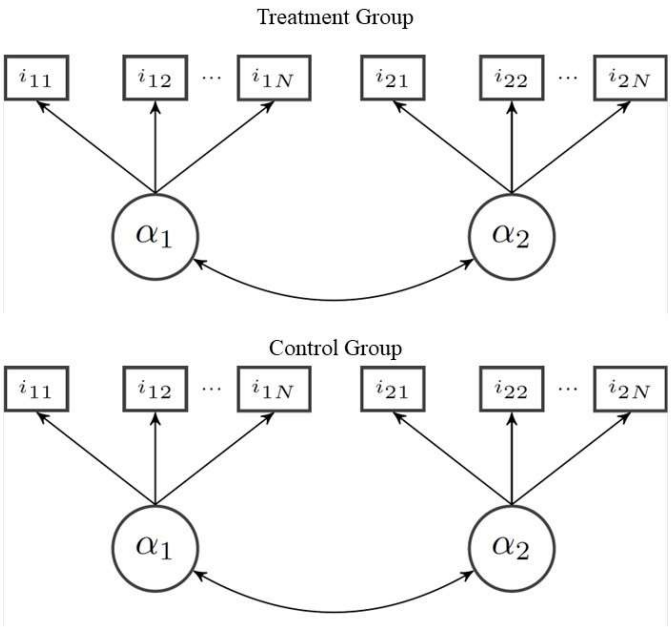

Figure 2. *Multi-timepoint Multigroup Measurement Model*

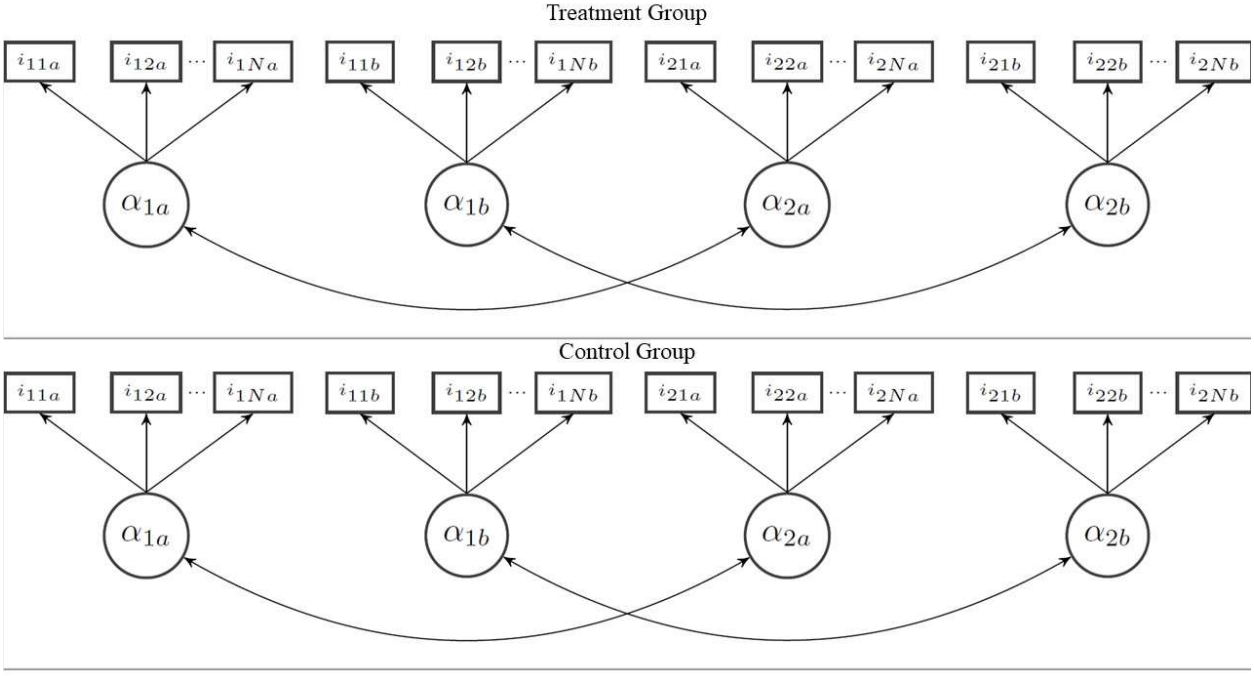

Figure 3. *Multi-timepoint Multigroup Measurement Model with Subgroups*

IRT Scoring Approaches

For the single timepoint calibration approach, we used two scoring approaches: MLE and EAP. For the Multigroup MIRT we used only EAP. MLE is not recommended for use with multi-dimensional models because it does not use information from the population distribution, which can cause convergence issues in score estimation, and undefined standard errors (Vector Psychometric Group, 2021). In combination, calibration and scoring approaches led to three distinct sets of scores that were compared to the original sum scoring approach that was used by Whitney & Candelaria (2017). Altogether, these combinations included: a univariate model calibrated with pooled timepoints scored with EAP, a univariate model with pooled timepoints scored with MLE, and a MIRT model calibrated with all timepoints scored with EAP. For Whitney and Candelaria's subgroup analysis, we also incorporated the SES and gender demographic subgroups into the MIRT model.

Estimating the Difference-in-Difference

For each scoring approach we estimated the Difference in Difference model using the specification originally used by Whitney and Candelaria (2017) reflected in the below equation:

$$Y_{ist} = \mu_s + \beta_1 Post_t + \beta_2(T_s x Post_t) + \boldsymbol{X}'_{it}\gamma + \varepsilon_{ist} \qquad (3)$$

where $Y_{ist}$is the socioemotional outcome of interest for student *i* in state *s* in year *t*, standardized within year *t*; $Post_t$ is a binary indicator variable that takes value 1 in the year 2003–2004 and is equal to 0 in the year 2001–2002; T_s is a binary indicator variable that takes value 1 if a student resides in a state that did not have prior consequential accountability and 0 otherwise; μ_s is a state-specific fixed effect; $\boldsymbol{X}_{it}$ is a vector of time-varying and time-invariant covariates; and

ε_{ist} is a mean-zero random error term. We included the model specification robustness checks the results of which are included in Appendix II but focused on the specification of the model that used state fixed effects and included the covariates but was not weighted for the total population analysis. For the subgroup analysis we used the same unweighted no covariate specification that Whitney and Candelaria (2017) used.

Evaluation Measures

To evaluate the impact of different scoring methods we focused on the treatment estimates produced. The original study used standardized sum scores putting them on the same scale as the scores produced by IRT methods. This allowed us to examine the shifts in point estimates of the treatment effect and standard errors in SDs and the impacts on interpretations that these shifts had.

Results

Treatment Effect Point Estimates

Overall, there was some, but not substantive, variation in the treatment effect point estimates produced by each scoring method. Sum scoring results matched the estimates from the original Whitney and Candelaria (2017) study exactly and the unidimensional scoring methods produced particularly similar point estimates.

All Students

Point estimates for the entire student sample over the various outcomes are shown in Figure 4. Across outcomes there were no perfectly consistent relationships between scoring methods, but the MGMT EAP method often produced an estimate that was larger in magnitude compared to the other methods. There was greater than a .02 SD difference between the smallest and largest

point estimates produced using the different scoring methods for five out of the ten outcomes with the Externalizing Behavior estimates the farthest apart at .04 SD. Generally, the univariate IRT scoring methods were close to each other for a given construct with an average difference of .008 SDs between the estimates over all the outcomes, indicating that the scoring approach had a minor impact on treatment estimates relative to the calibration approach.

The impact that scoring method did have likely reflects that the MGMT model better matches the structure of the data and the logic of the DiD design. The treatment and control groups in this study encapsulated students in states that had state-level accountability before NCLB and those that did not and the pre-post treatment timepoints were greater than a year apart. Using the MGMT model, which can incorporate into measurement differences by timepoint or group, is more in-line with the data-generating process, and therefore more likely to capture a true treatment effect than a scoring method that assumes that students in the treatment and control groups are exchangeable.

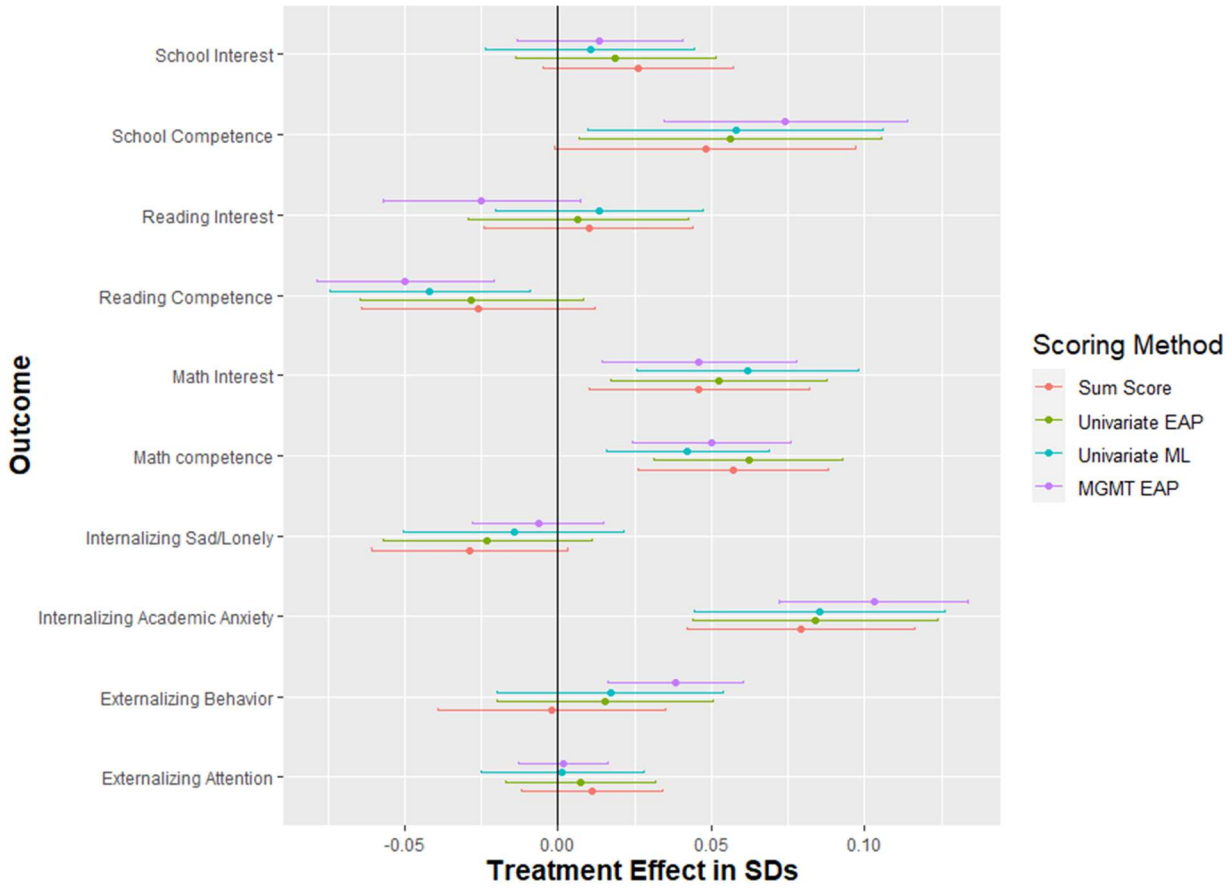

Figure 4. *Treatment Estimate with Standard Error Bars by Scoring Method (Total Population)*

Subgroups

..... Overall, there were more and larger differences between the treatment effects produced by each scoring method in the subgroup analysis as seen in Figure 5. The median difference between the smallest and largest point estimates across all scoring methods in the SES subgroup was .04 SDs and the maximum difference was .07 in School Interest Bottom 50% SES. Similarly, for the comparisons by sex, the median difference between the smallest and largest point estimates across all scoring methods was .035 SDs. As seen in Figures 5 and 6, these differences often widened the gaps between the subgroups of each outcome. The average gap

was .049 SDs between SES subgroups for the MGMT EAP that took into account SES and .039 SDs for sum scoring and .034 for regular MGMT EAP. For the sex subgroups, the average gap was .058 SDs for the MGMT EAP that took into account sex and .025 SDs for sum scoring and .023 for regular MGMT EAP. Notably, for externalizing behavior with sex groups the scoring methods resulted in large differences as seen in Figure 6. We discuss possible explanations for this in more detail in the discussion.

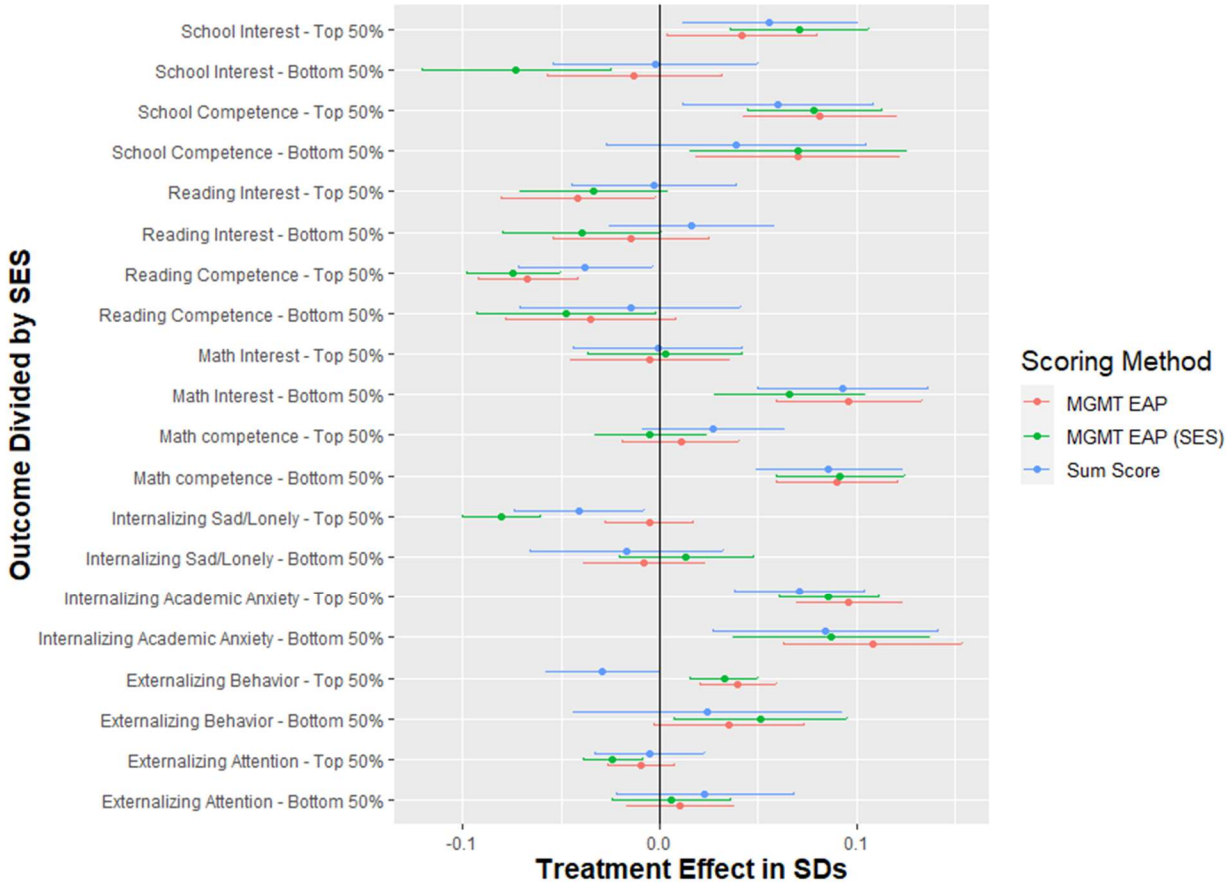

Figure 5. *Treatment Estimate with Standard Error Bars by Scoring Method (SES Subgroups)*

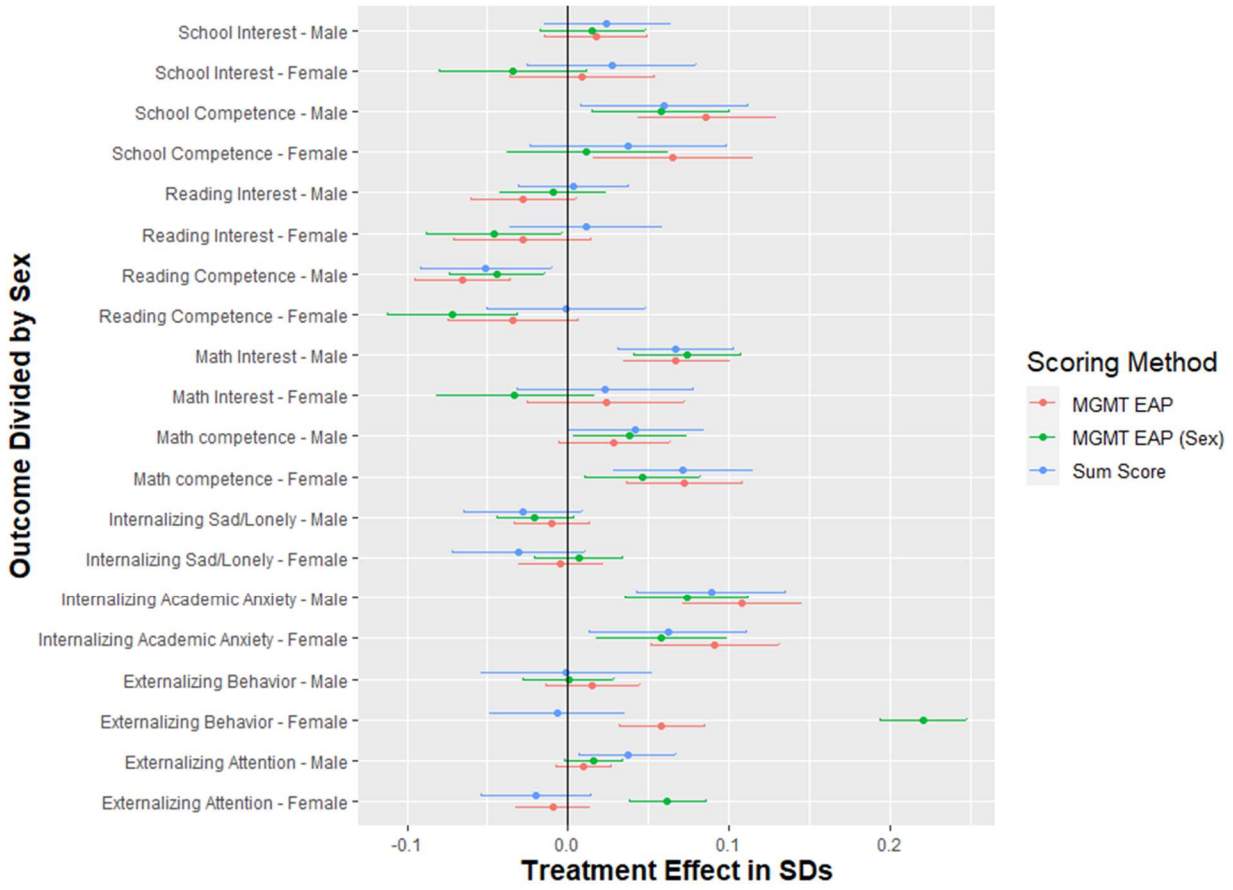

Figure 6. *Treatment Estimate with Standard Error Bar by Scoring Method (Sex Subgroups)*

Power and Significance

One of the most important aspects of a study is the ability to detect effects. Even for well-designed educational interventions most effect sizes are below .2 (Cheung & Slavin, 2016; Kraft, 2020) so studies need to be able to detect effects smaller than that. The scoring method influenced the ability to detect small effects through the standard errors (standard errors can be seen in Appendix I). In particular, the MGMT EAP scoring method consistently shrank standard errors. On average across outcomes the standard errors the MGMT EAP method produced were .0075 SDs smaller than the ones produced by the sum score when considering all students and

.01 SDs smaller than those produced by sum scores in the subgroups. These results reflected the shrinkage towards individual group/timepoints means of the MGMT EAP method.

Accordingly, the smaller standard errors and differential magnitude of the treatment effect estimates resulted in shifts in the p-values for many of the outcomes. When considering all students, the change in p-values associated with the MGMT EAP was enough to shift the significance levels for 3/10 outcomes. In the SES subgroups 8/20 had significance level shifts and in the sex subgroups 4/20 did. Some of these shifts were drastic. For example, the estimate of how much NCLB increased academic anxiety in all students had a p-value of .04 originally but using MGMT EAP produced a p-value of .003. Overall, the use of MGMT EAP increased the number of effects identified as significant. For all students, four of the ten constructs produced significant treatment effects using MGMT EAP compared to only two when using sum scores.

Discussion

Measurement decisions have been shown to affect the results of studies from those that use growth modeling (Kuhfeld & Soland, 2020) to those that use regression discontinuity designs (Soland, Johnson, et al., 2022). However, most of this work has been focused on simulation and does not apply specifically to the DiD design. This leaves the question of what practical implications measurement decisions have for DiD designs. We began to answer this question by showcasing how treatment estimates change in a DiD study that used several short survey scales and used a subgroup analysis to examine differential treatment effects.

The results showed that the scoring method can affect the significance level for all students even if it only shifted the point estimates slightly. Using a scoring method that incorporated the

muti-group and multi-timepoint nature of the DiD design tended to produce larger treatment effect estimates that were more often significant, a result in line with findings from Monte Carlo studies (e.g., Soland, Kuhfeld, & Edwards, 2022).

For example, consider the treatment effect of NCLB on academic anxiety The original study used a sum scoring approach and recovered a treatment effect that ranged from .076 to .145 SDs depending on the model specification and only had 4/8 specifications significant at a .1 alpha level. In contrast, using the MGMT scoring method in our study we recovered a treatment effect ranging from .1 to .159 SDs and every specification was significant at a .1 alpha level with the main specification producing a p-value of .003. In short, the scoring approach tended to have a substantive effect on statistical power. Some of this increase in statistical power comes from the reduction in the standard error driven by how MGMT shrinks scores towards the mean of each group. This increases the likelihood of false positives. On the other hand, alignment between study design and measurement decisions should result in more accurate point estimates and allow researchers to make more valid inferences including when there is a null or minimal treatment effect.

The importance of measurement decisions was reinforced by results of the subgroup analysis. When investigating the potential differential effect of a policy or treatment it is sensible to also incorporate the demographic trait that the researcher is investigating into the measurement model (Curran et al., 2016). This helps distinguish true treatment effects from those caused by measurement misspecification. In our case, incorporating demographic traits into MGMT scoring revealed several strong effects that were not apparent in the original study, while also strengthening or weakening the evidence for other findings. Particularly notable was the increase in externalizing behavior for female students that was estimated as -.007 in the original study, but

was estimated as .221 when using the MGMT EAP method that took into account sex. This effect was twice as large as the increase in academic anxiety (.079) , the main effect from the aggregate student analysis in the original study.

Crucially, externalizing behavior was a short measure and there were theoretical reasons to believe that scoring should be different by sex. The questions that made up the externalizing behavior measure focused on behavior that was disruptive. These questions are theoretically not likely to capture externalizing behavior equally well for male and female students because each sex tends to have different disruptive behaviors. Male students are much more likely to get into and punished for fighting (Lo & Cartledge, 2007; Mendez & Knoff, 2003; Monge et al., 2000) while girls can be criticized for talking loudly and boisterously (Koonce, 2012; Morris, 2007; Murphy et al., 2013). Thus, our results clearly outline how important scoring can be when conducting analysis of subgroups.

Overall, using scoring procedures that were aligned with the logic of the DiD and subgroup analysis seemed to give results that would have strengthened the original study. If the authors used MGMT EAP scoring the findings would have helped the authors make stronger claims about the impact of NCLB and revealed other differential treatment effects that should be investigated. While the results of our study indicate that the scoring approach has some influence on treatment effect estimates, especially the significance of results, we are unable to verify whether using the MGMT EAP method was more accurate. However, prior simulation research has shown that the MGMT EAP is better at recovering true treatment when data is collected at multiple timepoints (Kuhfeld & Soland, 2020; Soland, 2022, 2022). This combined with the theoretical alignment between MGMT EAP and the logic DiD design suggests that it should be used instead of sum scoring. While there is no gurantee any scoring method is the correct one,

MGMT EAP aligns more closely than the alternate scoring methods that we examined. Even if it is not the correct or "true" measurement model it is likely to perform better than one that is likely to be highly misspecified.

Limitations and Future Directions

Our study has a few limitations that bear mention. The primary limitation is that this is a re-analysis of a single study, so the results are not necessarily generalizable. In particular, the short lengths of the measures are likely an important factor that enhances the impact of using different scoring methods and may not apply to longer survey scales. Conversely, the relatively simple DiD design used means that there may be additional complications in using scoring methods with more complex DiD designs. More research that uses different measures and explores more complicated DiD designs is needed.

Another important consideration is that this study assumes that any underlying differences in measurement is caused by shifts in how participants understand and respond to survey questions relative to each other. Measurement differences could also reflect threats to a DiD design's validity that cannot be adjusted for by using different scoring approaches. For example, in Whitney and Candelaria's study if there was an incentive to show increases in reading interest in the treatment group after the NCLB is passed this might result in an over-reporting of reading interest in the treatment group. While this would be a measurement change that biases the treatment effect, it generally would not affect the loadings of items and would be undetectable using statistical procedures. Measurement changes like that one should be treated as a violation of the DiD identification assumptions and must be detected using tools like interviews rather than statistical procedures.

Conclusion

Our study shows that scoring can matter and can affect the conclusions of an empirical analysis with policy implications. In line with mathematical derivations and simulation research on survey scoring generally (Kuhfeld & Soland, 2020; Soland, 2021, 2022; Soland, Kuhfeld, et al., 2022) and specific to the DiD design (Chapter 1) some of the largest differences in estimates produced by scoring were in the shortest measures. When sample sizes are large enough, our results support an argument that researchers should seek to incorporate what they know about how the data are collected into the scoring model. For subgroup analyses specifically, including relevant demographic factors in the measurement model can help distinguish between differences in measurement and differential treatment effects across groups.

Chapter 2 References

Atkins-Burnett, S., & Meisels, S. J. (2001). Measures of Socio-Emotional Development in Middle Childhood. Working Paper No. 2001-03. *National Center for Education Statistics*.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, *16*(1), 87–96.

Bauer, D., & Curran, P. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 3–38.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Callaway, B., & Sant'Anna, P. H. (2020). Difference-in-differences with multiple time periods. *Journal of Econometrics*.

Caniglia, E. C., & Murray, E. J. (2020). Difference-in-Difference in the Time of Cholera: A Gentle Introduction for Epidemiologists. *Current Epidemiology Reports*, *7*(4), 203–211. https://doi.org/10.1007/s40471-020-00245-2

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Curran, P. J., Cole, V., Bauer, D. J., Hussong, A. M., & Gottfredson, N. (2016). Improving factor score estimation through the use of observed background characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(6), 827–844.

Delaney, J. A., & Kearney, T. D. (2015). Guaranteed tuition policies and state general appropriations for higher education: A difference-in-difference analysis. *Journal of Education Finance*, *40*(4), 359–390.

Eklund, K., Kilpatrick, K. D., Kilgus, S. P., & Haider, A. (2018). A systematic review of state-level social–emotional learning standards: Implications for practice and research. *School Psychology Review*, *47*(3), 316–326.

Fredricks, J. A., Hofkens, T., Wang, M.-T., Mortenson, E., & Scott, P. (2018). Supporting girls' and boys' engagement in math and science learning: A mixed methods study. *Journal of Research in Science Teaching*, *55*(2), 271–298.

Gehlbach, H., & Hough, H. J. (2018). Measuring Social Emotional Learning through Student Surveys in the CORE Districts: A Pragmatic Approach to Validity and Reliability. *Policy Analysis for California Education, PACE*.

Gorter, R., Fox, J.-P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology*, *79*, 140–149.

Gorter, R., Fox, J.-P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, *15*(1), 1–12.

Grant, S., Hamilton, L. S., Wrabel, S. L., Gomez, C. J., Whitaker, A. A., Leschitz, J. T., Unlu, F., Chavez-Herrerias, E. R., Baker, G., Barrett, M., Harris, M. A., & Ramos, A. (2017). *Social and Emotional Learning Interventions Under the Every Student Succeeds Act: Evidence Review*. RAND Corporation. https://www.rand.org/pubs/research_reports/RR2133.html

Koonce, J. B. (2012). Oh, those loud Black girls!": A phenomenological study of Black girls talking with an attitude. *Journal of Language and Literacy Education*, *8*(2), 26–46.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kuhfeld, M., & Soland, J. (2020). Avoiding bias from sum scores in growth estimates: An examination of IRT-based approaches to scoring longitudinal survey responses. *Psychological Methods*. https://doi.org/10.1037/met0000367

Lo, Y., & Cartledge, G. (2007). Office disciplinary referrals in an urban elementary school. *Multicultural Learning and Teaching*, *2*(1).

Mahoney, J. L., Weissberg, R. P., Greenberg, M. T., Dusenbury, L., Jagers, R. J., Niemi, K., Schlinger, M., Schlund, J., Shriver, T. P., & VanAusdal, K. (2020). Systemic social and emotional learning: Promoting educational success for all preschool to high school students. *American Psychologist*.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.

Mendez, L. M. R., & Knoff, H. M. (2003). Who gets suspended from school and why: A demographic analysis of schools and disciplinary infractions in a large school district. *Education and Treatment of Children*, 30–51.

Monge, E., Massie, J., Larson, K., & Sarvela, P. (2000). Predictors of fighting among rural elementary school students. *Journal of Health Education*, *31*(2), 69–73.

Morris, E. W. (2007). "Ladies" or "loudies"? Perceptions and experiences of Black girls in classrooms. *Youth & Society*, *38*(4), 490–515.

Murphy, A. S., Acosta, M. A., & Kennedy-Lewis, B. L. (2013). “I’m not running around with my pants sagging, so how am I not acting like a lady?”: Intersections of race and gender in the experiences of female middle school troublemakers. *The Urban Review*, *45*(5), 586–610.

Pollack, J. M., Najarian, M., Rock, D. A., & Atkins-Burnett, S. (2005). Early Childhood Longitudinal Study, Kindergarten Class of 1998? 99 (ECLS-K). Psychometric Report for the Fifth Grade. NCES 2006? 036. *National Center for Education Statistics*.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*.

Soland, J. (2021). Is measurement noninvariance a threat to inferences drawn from randomized control trials? Evidence from empirical and simulation studies. *Applied Psychological Measurement*, 01466216211013102.

Soland, J. (2022). Evidence That Selecting an Appropriate Item Response Theory–Based Approach to Scoring Surveys Can Help Avoid Biased Treatment Effect Estimates. *Educational and Psychological Measurement*, 00131644211007551.

Soland, J., Johnson, A., & Talbert, E. (2022). Regression Discontinuity Designs in a Latent Variable Framework. *Psychological Methods*.

Soland, J., Kuhfeld, M., & Edwards, K. (2022). How survey scoring decisions can influence your study’s results: A trip through the IRT looking glass. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000506

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2013). The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation. *MDRC*.

Talbert, E., Hofkens, T., & Wang, M.-T. (2019). Does student-centered instruction engage students differently? The moderation effect of student ethnicity. *The Journal of Educational Research*, *112*(3), 327–341.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., & Najarian, M. (2009). Early Childhood Longitudinal Study, Kindergarten Class of 1998-99 (ECLS-K): Combined User's Manual for the ECLS-K Eighth-Grade and K-8 Full Sample Data Files and Electronic Codebooks. NCES 2009-004. *National Center for Education Statistics*.

Whitney, C. R., & Candelaria, C. A. (2017). The effects of No Child Left Behind on children's socioemotional outcomes. *AERA Open*, *3*(3), 2332858417726324.

Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, *39*.

Chapter 2 Appendix I: Main Estimate Tables

Table A1.1 Total Population Coefficients and Standard Errors by Outcome (Unweighted State Fixed Effects Model with covariates)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	0.011	(0.023)	0.007	(0.024)	0.001	(0.027)	0.002	(0.015)
Externalizing Behavior	-0.002	(0.037)	0.015	(0.035)	0.017	(0.037)	0.038	(0.022)*
Internalizing Sad/Lonely	-0.029	(0.032)	-0.023	(0.034)	-0.015	(0.036)	-0.007	(0.021)
Internalizing Academic Anxiety	0.079	(0.037)**	0.084	(0.04)**	0.085	(0.041)**	0.103	(0.031)***
Math Interest	0.046	(0.036)	0.052	(0.035)	0.062	(0.036)	0.046	(0.032)
Math Competence	0.057	(0.031)*	0.062	(0.031)*	0.042	(0.027)	0.050	(0.026)*
Reading Interest	0.010	(0.034)	0.007	(0.036)	0.013	(0.034)	-0.025	(0.032)
Reading Competence	-0.026	(0.038)	-0.028	(0.036)	-0.042	(0.033)	-0.050	(0.029)
School Interest	0.026	(0.031)	0.019	(0.033)	0.011	(0.034)	0.014	(0.027)
School Competence	0.048	(0.049)	0.056	(0.049)	0.058	(0.048)	0.074	(0.04)*

Note. * p<.1, **p< .01, ***p<.001

Table A1.2 SES Subgroup Coefficients and Standard Errors by Outcome (Unweighted No Covariate Model)

Outcome	Subgroup	Sum Scoring		MGMT EAP		MGMT EAP (SES)	
Externalizing Attention	Bottom SES	0.023	(0.045)	0.010	(0.027)	0.006	(0.03)
Externalizing Attention	Top SES	-0.005	(0.028)	-0.010	(0.017)	-0.024	(0.015)
Externalizing Behavior	Bottom SES	0.024	(0.068)	0.035	(0.038)	0.051	(0.044)
Externalizing Behavior	Top SES	-0.029	(0.029)	0.040	(0.019)**	0.033	(0.017)
Internalizing Sad/Lonely	Bottom SES	-0.017	(0.049)	-0.008	(0.031)	-0.013	(0.034)
Internalizing Sad/Lonely	Top SES	-0.041	(0.033)	-0.005	(0.022)	-0.081	(0.02)***
Internalizing Academic Anxiety	Bottom SES	0.084	(0.057)	0.108	(0.045)**	0.087	(0.05)*
Internalizing Academic Anxiety	Top SES	0.071	(0.033)**	0.096	(0.027)***	0.086	(0.025)***
Math Interest	Bottom SES	0.093	(0.043)**	0.096	(0.037)**	0.066	(0.038)*
Math Interest	Top SES	-0.001	(0.043)	-0.005	(0.04)	-0.003	(0.039)
Math Competence	Bottom SES	0.086	(0.037)**	0.090	(0.031)***	0.092	(0.032)***
Math Competence	Top SES	-0.027	(0.036)	-0.011	(0.03)	-0.005	(0.028)

Reading Interest	Bottom SES	0.016	(0.042)	-0.014	(0.04)	-0.039	(0.04)
Reading Interest	Top SES	-0.003	(0.042)	-0.042	(0.039)	-0.034	(0.038)
Reading Competence	Bottom SES	-0.015	(0.056)	-0.035	(0.043)	-0.047	(0.046)
Reading Competence	Top SES	-0.038	(0.034)	-0.067	(0.025)**	-0.074	(0.024)***
School Interest	Bottom SES	-0.002	(0.052)	-0.013	(0.044)	-0.073	(0.048)
School Interest	Top SES	0.056	(0.044)	0.042	(0.038)	0.071	(0.035)*
School Competence	Bottom SES	0.039	(0.066)	0.070	(0.051)	0.070	(0.055)
School Competence	Top SES	0.06	(0.048)	0.081	(0.039)**	0.079	(0.034)**

Note. * p<.1, **p< .01, ***p<.001

Table A1.3 Sex Subgroup Coefficients and Standard Errors by Outcome (Unweighted No Covariate Model)

Outcome	Subgroup	Sum Scoring		MGMT EAP		MGMT EAP (Sex)	
Externalizing Attention	Male	0.037	(0.03)	0.010	(0.017)	0.016	(0.018)
Externalizing Attention	Female	-0.02	(0.034)	-0.010	(0.023)	0.062	(0.024)**
Externalizing Behavior	Male	-0.001	(0.053)	0.015	(0.029)	0.000	(0.028)
Externalizing Behavior	Female	-0.007	(0.042)	0.058	(0.026)	0.221	(0.027)***
Internalizing Sad/Lonely	Male	-0.028	(0.037)	-0.010	(0.023)	-0.021	(0.024)
Internalizing Sad/Lonely	Female	-0.031	(0.041)	-0.005	(0.026)	0.007	(0.027)
Internalizing Academic Anxiety	Male	0.089	(0.046)*	0.108	(0.037)***	0.074	(0.038)*
Internalizing Academic Anxiety	Female	0.062	(0.049)	0.091	(0.04)**	0.058	(0.04)
Math Interest	Male	0.067	(0.036)	0.067	(0.032)**	0.074	(0.033)*
Math Interest	Female	0.023	(0.055)	0.024	(0.049)	-0.033	(0.048)
Math Competence	Male	0.042	(0.042)	0.029	(0.034)	0.038	(0.035)
Math Competence	Female	0.071	(0.043)*	0.072	(0.036)*	0.046	(0.036)

Reading Interest	Male	0.003	(0.034)	-0.028	(0.032)	-0.009	(0.033)
Reading Interest	Female	0.011	(0.047)	-0.028	(0.042)	-0.046	(0.042)
Reading Competence	Male	-0.051	(0.041)	-0.065	(0.03)	-0.044	(0.03)
Reading Competence	Female	-0.001	(0.049)	-0.034	(0.04)*	-0.072	(0.04)*
School Interest	Male	0.024	(0.039)	0.017	(0.032)	0.015	(0.032)
School Interest	Female	0.027	(0.052)	0.009	(0.045)	-0.034	(0.046)
School Competence	Male	0.06	(0.052)	0.086	(0.043)*	0.058	(0.043)
School Competence	Female	0.037	(0.061)	0.065	(0.05)	0.012	(0.05)

Note. * p<.1, **p< .01, ***p<.001

Appendix II: Robustness Check Results

Table A2.1 Total Population Coefficients and Standard Errors by Outcome (Unweighted State Fixed Effects No Covariates Model)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	0.009	*(0.024)*	0.005	*(0.034)*	-0.001	*(0.027)*	-0.001	*(0.015)*
Externalizing Behavior	-0.004	*(0.039)*	0.013	*(0.037)*	0.016	*(0.038)*	0.036	*(0.024)*
Internalizing Sad/Lonely	-0.029	*(0.032)*	-0.023	*(0.034)*	-0.015	*(0.036)*	-0.007	*(0.021)*
Internalizing Academic Anxiety	0.076	*(0.037)***	0.081	*(0.04)*	0.083	*(0.041)**	0.100	*(0.031)****
Math Interest	0.044	*(0.035)*	0.051	*(0.034)*	0.060	*(0.036)*	0.045	*(0.031)*
Math Competence	0.055	*(0.032)**	0.059	*(0.031)*	0.041	*(0.027)*	0.049	*(0.026)**
Reading Interest	0.007	*(0.036)*	0.004	*(0.037)*	0.011	*(0.035)*	-0.028	*(0.034)*
Reading Competence	-0.026	*(0.037)*	-0.028	*(0.036)*	-0.042	*(0.032)*	-0.049	*(0.028)*
School Interest	0.025	*(0.031)*	0.018	*(0.033)*	0.010	*(0.035)*	0.013	*(0.027)*
School Competence	0.048	*(0.049)*	0.056	*(0.049)*	0.058	*(0.049)*	0.075	*(0.04)**

Note. * p<.1, **p< .01, ***p<.001

Table A2.3 Total Population Coefficients and Standard Errors by Outcome (Unweighted Child Fixed Effects No Covariates Model)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	0.008	*(0.033)*	0.006	*(0.034)*	0.000	*(0.037)*	0.002	*(0.02)*
Externalizing Behavior	0.005	*(0.053)*	0.022	*(0.05)*	0.022	*(0.053)*	0.044	*(0.03)*
Internalizing Sad/Lonely	-0.024	*(0.043)*	-0.019	*(0.045)*	-0.010	*(0.049)*	0.001	*(0.025)*
Internalizing Academic Anxiety	0.079	*(0.053)*	0.083	*(0.056)*	0.085	*(0.058)*	0.100	*(0.042)***
Math Interest	0.047	*(0.05)*	0.052	*(0.049)*	0.060	*(0.051)*	0.045	*(0.043)*
Math Competence	0.061	*(0.044)*	0.064	*(0.043)*	0.044	*(0.038)*	0.051	*(0.036)*
Reading Interest	0.015	*(0.05)*	0.013	*(0.051)*	0.022	*(0.048)*	-0.018	*(0.043)*
Reading Competence	-0.020	*(0.05)*	-0.022	*(0.049)*	-0.038	*(0.044)*	-0.044	*(0.037)*
School Interest	0.025	*(0.042)*	0.019	*(0.045)*	0.012	*(0.047)*	0.013	*(0.037)*
School Competence	0.049	*(0.069)*	0.058	*(0.069)*	0.060	*(0.068)*	0.075	*(0.055)*

Note. * p<.1, **p< .01, ***p<.001

Table A2.4 Total Population Coefficients and Standard Errors by Outcome (Weighted State Fixed Effects Model No covariates)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	-0.010	*(0.05)*	-0.023	*(0.05)*	-0.010	*(0.053)*	-0.019	*(0.032)*
Externalizing Behavior	-0.058	*(0.064)*	-0.061	*(0.064)*	-0.077	*(0.064)*	-0.017	*(0.044)*
Internalizing Sad/Lonely	-0.055	*(0.057)*	-0.044	*(0.055)*	-0.023	*(0.052)*	-0.024	*(0.036)*
Internalizing Academic Anxiety	0.132	*(0.076)**	0.154	*(0.074)***	0.159	*(0.073)***	0.153	*(0.058)***
Math Interest	0.066	*(0.047)*	0.054	*(0.047)*	0.020	*(0.055)*	0.050	*(0.043)*
Math Competence	0.068	*(0.039)**	0.067	*(0.039)**	0.040	*(0.044)*	0.059	*(0.034)**
Reading Interest	0.013	*(0.056)*	0.002	*(0.061)*	0.001	*(0.059)*	-0.023	*(0.055)*
Reading Competence	0.031	*(0.066)*	0.026	*(0.063)*	0.000	*(0.055)*	-0.001	*(0.051)*
School Interest	0.028	*(0.047)*	0.021	*(0.049)*	0.009	*(0.05)*	0.019	*(0.042)*
School Competence	0.043	*(0.057)*	0.059	*(0.058)*	0.060	*(0.058)*	0.082	*(0.047)**

Note. * p<.1, **p< .01, ***p<.001

Table A2.5 Total Population Coefficients and Standard Errors by Outcome (Weighted State Fixed Effects Model with covariates)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	-0.004	*(0.049)*	-0.018	*(0.048)*	-0.004	*(0.051)*	-0.014	*(0.029)*
Externalizing Behavior	-0.049	*(0.062)*	-0.053	*(0.062)*	-0.070	*(0.063)*	-0.009	*(0.043)*
Internalizing Sad/Lonely	-0.051	*(0.055)*	-0.040	*(0.054)*	-0.018	*(0.051)*	-0.020	*(0.034)*
Internalizing Academic Anxiety	0.138	*(0.076)**	0.159	*(0.075)***	0.163	*(0.074)***	0.157	*(0.059)**
Math Interest	0.065	*(0.046)*	0.052	*(0.047)*	0.019	*(0.055)*	0.048	*(0.044)*
Math Competence	0.070	*(0.039)**	0.068	*(0.038)**	0.042	*(0.043)*	0.060	*(0.034)**
Reading Interest	0.015	*(0.055)*	0.003	*(0.06)*	0.002	*(0.059)*	-0.022	*(0.054)*
Reading Competence	0.029	*(0.066)*	0.024	*(0.063)*	-0.002	*(0.054)*	-0.003	*(0.051)*
School Interest	0.027	*(0.046)*	0.020	*(0.048)*	0.007	*(0.049)*	0.018	*(0.041)*
School Competence	0.045	*(0.056)*	0.060	*(0.057)*	0.062	*(0.057)*	0.083	*(0.046)**

Note. * p<.1, **p< .01, ***p<.001

Table A2.6 Total Population Coefficients and Standard Errors by Outcome (Weighted Child Fixed Effects with Covariates Model)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	-0.001	*(0.068)*	-0.012	*(0.066)*	0.004	*(0.07)*	-0.006	*(0.037)*
Externalizing Behavior	-0.026	*(0.081)*	-0.030	*(0.081)*	-0.048	*(0.082)*	0.013	*(0.049)*
Internalizing Sad/Lonely	-0.043	*(0.077)*	-0.029	*(0.073)*	-0.006	*(0.068)*	-0.011	*(0.042)*
Internalizing Academic Anxiety	0.145	*(0.103)*	0.163	*(0.103)*	0.168	*(0.102)*	0.159	*(0.079)**
Math Interest	0.060	*(0.071)*	0.047	*(0.072)*	0.017	*(0.082)*	0.042	*(0.066)*
Math Competence	0.057	*(0.06)*	0.054	*(0.06)*	0.026	*(0.066)*	0.044	*(0.052)*
Reading Interest	0.016	*(0.08)*	0.007	*(0.083)*	0.011	*(0.079)*	-0.018	*(0.072)*
Reading Competence	0.026	*(0.085)*	0.018	*(0.083)*	-0.015	*(0.075)*	-0.008	*(0.067)*
School Interest	0.021	*(0.066)*	0.015	*(0.067)*	0.008	*(0.066)*	0.012	*(0.056)*
School Competence	0.029	*(0.079)*	0.045	*(0.079)*	0.049	*(0.079)*	0.065	*(0.063)*

Note. * p<.1, **p< .01, ***p<.001

Table A2.7 Total Population Coefficients and Standard Errors by Outcome (Weighted Child Fixed Effects No Covariates Model)

Outcome	Sum Scoring		Univariate EAP		Univariate ML		MGMT EAP	
Externalizing Attention	-0.007	*(0.068)*	-0.017	*(0.066)*	-0.001	*(0.07)*	-0.010	*(0.037)*
Externalizing Behavior	-0.030	*(0.081)*	-0.035	*(0.08)*	-0.053	*(0.081)*	0.010	*(0.049)*
Internalizing Sad/Lonely	-0.044	*(0.078)*	-0.032	*(0.075)*	-0.009	*(0.07)*	-0.010	*(0.043)*
Internalizing Academic Anxiety	0.141	*(0.106)*	0.160	*(0.106)*	0.165	*(0.104)*	0.157	*(0.081)**
Math Interest	0.063	*(0.068)*	0.051	*(0.069)*	0.019	*(0.081)*	0.045	*(0.063)*
Math Competence	0.061	*(0.058)*	0.057	*(0.058)*	0.027	*(0.065)*	0.046	*(0.051)*
Reading Interest	0.015	*(0.078)*	0.007	*(0.08)*	0.010	*(0.076)*	-0.018	*(0.071)*
Reading Competence	0.033	*(0.088)*	0.026	*(0.086)*	-0.005	*(0.078)*	-0.002	*(0.069)*
School Interest	0.024	*(0.065)*	0.019	*(0.067)*	0.011	*(0.067)*	0.015	*(0.056)*
School Competence	0.030	*(0.081)*	0.046	*(0.082)*	0.048	*(0.083)*	0.066	*(0.066)*

Note. * p<.1, **p< .01, ***p<.001

Chapter 3: Combining Multi-Rater Observer Protocols with Regression Discontinuity Design for Unbiased Causal Effect Estimation

Chapter 3 Abstract

Regression Discontinuity (RD) designs are a valuable tool for estimating causal effects in educational and behavioral research when randomized control trials (RCTs) are not feasible. In education, they are commonly used to examine outcomes from multi-item measures, such as survey scales, teacher/student observational protocols, and tests. However, observational protocols that rely on multiple raters present a unique challenge, as multiple ratings must be combined and previous research has highlighted the limitations of using simplistic measurement models like those implicit to sum scores. To address this challenge, we propose a new model, the RD Multi-rater model, which incorporates information from multiple raters in a structural equation modeling (SEM) framework. Through simulation, we investigate the feasibility of the model and assess the reduction in bias and Type II errors in comparison to other methods for handling multiple ratings. We find sample size is the main limiting factor but even using a suboptimal sample size the RD Multi-rater model still performs better than more simplistic measurement models.

Keywords: Regression discontinuity, observational protocols, structural equation modeling, multi-rater

Combining Multi-Rater Observer Protocols with Regression Discontinuity Design for Unbiased Causal Effect Estimation

Regression Discontinuity (RD) designs can be used to estimate unbiased causal effects in educational and behavioral research when randomized control trials (RCTs) are not feasible (Bloom, 2012; Hahn et al., 2001; Lee & Lemieux, 2010). In an RD, units are assigned to treatment based on a cutoff score on one or more variables. If one assumes units close to either side of the cutoff only differ by virtue of measurement error on the assessment being used to make treatment assignments (Berk et al., 2010; Black et al., 2005; Shadish et al., 2011), then those units are ignorably assigned to treatment. When this assumption is met, RDs produce results comparable to those from RCTs (Aiken et al., 1998; Berk et al., 2010; Black et al., 2005; Shadish et al., 2011).

In fields like education, the outcomes RDs are used to assess typically come from multi-item measures including survey scales, teacher/student observational protocols, and tests. For non-achievement measures, scores are often generated using a simple sum score model (Bauer & Curran, 2015) even though sum scores rely on large, oftentimes untenable assumptions (McNeish & Wolf, 2020). Flake et al. (2017) quantified the wide use of sum scores in psychology by reporting that only 21% (37 out of 177) of the studies they reviewed used a latent variable model rather than a sum score and a mere 2% of author-developed scales reported any evidence of internal structure (3 out of 124). Thus, in education and psychology, when standardized test scores are not the dependent variable, results are often based on scores from measurement models that are simplistic and make unjustifiable assumptions.

In an RD context, Soland et al. (2022) revealed the potential benefits of better incorporating measurement models for latent variables into the design rather than using sum scores. They

showed using a latent variable model rather than sum scores avoids biased treatment effect estimates that can result when the assumptions of sum scoring are not met under realistic conditions. However, these benefits were explored mostly in the context of short survey measures. There are other multi-item measures that may be used in RD designs. In particular, observational protocols are increasingly likely to be included in educational administrative datasets and RDs have already been used in the teacher evaluation (Dee & Wyckoff, 2015) , classroom quality improvement (Bassok et al., 2019), and child development (Hutcheon et al., 2020) contexts. These observational protocols have documented limitations. Responses to observational protocols rely on rater knowledge of the target, which may be limited, and other factors like who does the rating and the time of the rating can introduce extraneous sources of variation. For example, Ho and Kane (2013) when examining a teacher observation protocol based on Charlotte Danielson's Framework for Teaching (Danielson, 2014), when administrators rather than a teacher's peers conducted the rating the variance of the construct was 50% higher. While a construct might legitimately vary across factors, researchers are often interested in the stable component of the construct. Therefore, to compensate for these limitations multiple raters can be utilized, but those ratings need to be combined in a psychometrically defensible way. Failure to use psychometrically defensible aggregation methods can lead to challenges, such as a lack of optimal informant selection (Kraemer et al., 2003), and may result in no improvement in predictive validity compared to a single rater (van Dulmen & Egeland, 2011).

However, there are many psychometrically defensible models and choosing one requires its own considerations. For example, the Correlated Trait/Correlated Uniqueness model (CTCU; Marsh, 1989), an early model for multi-rater data, has been criticized for failing to separate

method variance from error variance, potentially leading to underestimation of the reliability of the measured variables (Eid, 2000; Lance et al., 2002; Pohl & Steyer, 2010).

To choose a model we considered the situations where multiple raters may be used in educational evaluation. Broadly speaking there are three scenarios that multiple raters might be used in education. In the first scenario, multiple highly trained raters rate a target under very similar conditions. An example of this scenario is when several raters grade an essay using a predetermined rubric or when trained researchers rate a recorded teaching session. Another scenario is where ratings happen under similar but much less controlled conditions. Live ratings of teachers or classes as well as when raters have variable training are all examples of this type of scenario. In the last scenario, researchers are interested in the different perspectives of the raters, such as that provided by teachers and parents. Both the raters and the situations they conduct their ratings intentionally vary and raters often are untrained or minimally trained. For example, to collect data for the NICHD Study of Early Child Care and Youth Development information was collected from both the teachers and parents of students (Network, 2006). Given the resource and other logistical constraints that accompany large number of teacher and student evaluations, the latter two varied rater situations are more likely in educational evaluation.

Accordingly, for this study, we identified the trifactor model proposed by Bauer et al. (2013) as a model that was suitable for these varied rater scenarios. The tri-factor model partitions variance between raters, the observation protocol items, and the target construct making it easy to understand and flexible enough to accommodate the varied rater scenarios where there are many sources of variability. Building upon Soland et al. (2022), we introduce the RD Multi-rater model, which integrates the SEM model from Soland et al. (2022) and Bauer et al.'s tri-factor

model. However, combining these two models requires estimation of at least six parameters per shared item as well as the structural components of the RD, a large number of parameters, and the tri-factor model has not been used extensively with causal evaluations. Thus, we begin by assessing the feasibility of our model, with a focus on the number of observations needed, and then evaluate its potential benefits.

In particular we are interested in two benefits. First, given a common rationale for using multiple raters is to improve accuracy of ratings, we are interested in how the RD Multi-rater's combination of raters affects bias and how this compares to using a single rater and misspecified models. Second, we are interested in how the use of multiple raters can affect variance and power given the estimate that in education an RD study needs between 9 and 17 times as many schools or students as an RCT to produce an impact with the same level of statistical precision (Deke & Dragoset, 2012). To achieve these objectives, we address two research questions that reflect critical considerations when employing this model:

1. Under what sample size conditions is it feasible for the RD Multi-rater model to recover treatment effects of various magnitudes?
2. To what extent does incorporating information from multiple raters in the RD Multi-rater model reduce bias and Type II errors in treatment effect estimates compared to using sum scores, ignoring multiple raters, or using single raters?

We explore these questions in two separate simulation studies, which will build on the results of Soland et al. (2022) by expanding the findings to an outcome assessed by multiple raters.

Background

In this section, we briefly describe the logic of the RD model we use before explaining how using a sum score versus a SEM for the dependent variable might generally affect power and

bias due to misspecifying the measurement model in ways common to sum scores. We first use a survey scale type latent variable as a simplified example before covering the additional complications of using observation protocols. Finally, we present a model that allows for the combination of observations from multiple raters in an RD design.

Regression Discontinuity Designs

RDs, like RCTs, estimate causal effects of a treatment by using a control group to approximate the counterfactual (Lee & Lemieux, 2010). In an RD, treatment is assigned based on one or more variables known as the running or assignment variable(s). This can be through a sharp cutoff score where the probability of treatment goes from 0 to 1 (sharp RDs) or a situation where the probability of receiving treatment is only strongly influenced by the cut score on the running variable(s) (fuzzy RDs). However, one limitation of the RD design is that, for the full sample to be used to estimate the casual effect at the discontinuity, the running variable must be properly controlled for. While one approach is to approximate the relationship between the running variable and the outcome using statistical information criteria or cross-validation to choose the degree of a higher order polynomial, Gelman and Imbens (2019) have shown that the use of higher order polynomials can lead to noisy estimates that are sensitive to the degree of the polynomial. For simplicity, in this paper, we discuss a parametric RD model that is applied to a bandwidth where a local linear or quadratic approximation is valid and cases outside the bandwidth are not used, in line with the approach of Hahn et al. (2001).

Turning to the specific RD model, let r_i be the rating score for person *i* (also referred to as the running variable or the forcing variable). This rating variable must be continuous or semi-continuous with discrete values, such as test scores with integer values (see Lee and Card [(2008)] and Kolesár and Rothe [(2018)] for using semi-continuous rating variables). Let r^* be

the cut score for this variable (also called the treatment threshold). Oftentimes, r_i is centered at r^* such that $r^* = 0$, which makes interpretation of regression coefficients more straightforward. A participant i's assigned treatment is represented by z_i, such that:

$$z_i = \mathrm{I}\{r_i \geq r^*\} = \begin{cases} 1 \; if \; r_i \geq r^* \\ 0 \; if \; r_i < r^* \end{cases}. \quad (1)$$

For now, we will assume that z_i is equal to t_i, the treatment the participant *actually* experiences (a sharp RD). If our observed score on the outcome of interest is denoted as obs_i, then a basic RD formulation would be.

$$obs_i = \beta_0 + \beta_1 r_i + \beta_2 z_i + \beta_3 z_i r_i + \epsilon_i \quad (2)$$

The interaction term, $\beta_3 z_i r_i$, allows for the slope to differ within the bandwidth near the cutoff. Further, quadratic terms can be added that are the same or different on either side of the cut score. As previously discussed, estimates are only unbiased if the functional form correctly models the relationship between treatment status and the outcome and does so on each side of the cut score. In our model the RD design estimates local average treatment effects using data with forcing variable values near the cut score where a local or quadratic proximation of the relationship between the outcome and the forcing variable is likely to hold. Results are thus influenced by the bandwidth used to estimate the treatment effect.

In the case of a fuzzy RD, z_i does not necessarily equal t_i. For instance, a policy might require students with a math score below a set threshold to receive extra instruction, but not all students below the cut score ultimately get the treatment. To address the issue of imperfect assignment, IVs are used to account for a potential correlation between treatment status and the outcome of interest. This adds another source of variability but corrects for imperfect assignment. In a fuzzy RD the assignment, z_i, serves as the instrument that affects the outcome

only through its effect on the actual treatment status, t_i. This exclusion of a direct causal link between the instrument and the outcome is called the "exclusion restriction" and helps address concerns about omitted variable bias. That is, we can estimate the indirect effect of z_i through t_i on the outcome, which does not suffer from selection bias, and use that indirect effect to produce an unbiased estimate of the direct effect of the treatment. Thus, the full RD specification involves submodels for both our observed outcome obs_i, the outcome of interest, and t_i, the actual treatment status. An example of these two-stage equations are below.

$$t_i = \alpha_0 + \alpha_1 r_i + \alpha_2 z_i + \alpha_3 z_i r_i + \mu_i \quad (3)$$

$$obs_i = \beta_0 + \beta_1 r_i + \beta_2 \hat{t}_i + \beta_3 z_i r_i + \epsilon_i \quad (4)$$

In the above equations, the treatment effect is being estimated based only on the variance in treatment status t_i that can be explained by assignment to treatment z_i, which is not affected by selection bias.

One can also express an IV model in the form of an SEM path diagram, depicted in Figure 1 (Murnane & Willett, 2010). In that figure, y_{1i} through y_{6i} are observed indicators of the construct of interest (e.g., survey item responses) and η_i is a latent variable underlying those observed indicators and is the outcome of interest. As the figure shows, there is a path from z_i to t_i, but no covariance between z_i and η_i otherwise—a visual representation of the exclusion restriction. Further, in this path diagram, the correlation in the error terms between η_i and t_i is expressed directly and estimated. One should note that, if z_i and related interaction terms were eliminated from the model, we would have the sharp RD design. Also, unlike in the two-stage approach, all constituent models in the RD are estimated simultaneously. We discuss the measurement and structural components of such models more below.

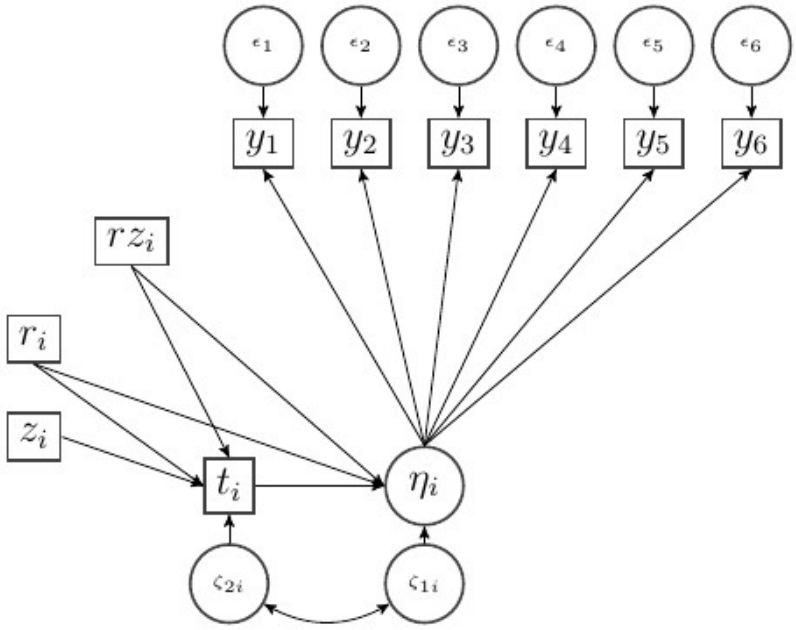

Figure 1. *Path diagram for an example fuzzy RD model.*

Measurement Bias Common to Sum Scores in RDs

A strong argument in favor of using an SEM to estimate an RD is avoiding bias common to sum scores and their assumptions. Misspecification of the measurement model used to score the dependent variable can lead to outright bias in parameter estimates, including treatment effect estimates (D. Bauer & Curran, 2015; Gorter et al., 2016; McNeish & Wolf, 2020; Soland et al., 2022). Such misspecification is especially likely when surveys or other measures are scored by simply summing or taking the average of the observed item responses (McNeish & Wolf, 2020). As McNeish and Wolf (2020) demonstrate, such sum score approaches are the equivalent of fitting a highly constrained measurement model that assumes (often wrongly) that items should be weighted equally (typically by constraining loadings equal to one, as discussed above) and that the error terms are equivalent across items (and typically equal to zero).

Use of sum scores to produce the dependent variable (or other related forms of model misspecification) could lead to biased treatment effect estimates in an RD. Let us consider the simple case of an SEM with a single latent variable measured by one or more observed indicators and a structural model in which treatment status for person i, t_i, is the only predictor (this could easily be extended from the RCT scenario we examine here to a sharp RD in which r_i and other relevant covariates are included). Our measurement model would be

$$\boldsymbol{y}_i = \boldsymbol{\nu} + \boldsymbol{\lambda}\eta_i + \boldsymbol{\epsilon}_i \quad (5)$$

Where $\boldsymbol{y}_i$ is an $N\ x\ 1$ vector of observed item responses for items $1\ldots N$, $\boldsymbol{\nu}$ is an $N\ x\ 1$ vector of intercepts, $\boldsymbol{\lambda}$ is an $N\ x\ 1$ vector of loadings, η_i is the single latent variable for person i, and $\boldsymbol{\epsilon}_i$ is an $N\ x\ 1$ vector of residuals with $VAR(\boldsymbol{\epsilon}_i) = \boldsymbol{\Theta} = diag(\theta_{11}, \theta_{22} \ldots \theta_{NN})$.

The structural model is

$$\eta_i = \alpha + \gamma t_i + \zeta_i \quad (6)$$

with true score variance ϕ and t_i treatment status.

For illustration, we will assume α_0 is zero (i.e., the mean η_i for the control group is zero) and $\boldsymbol{\nu}$ is a vector of zeros (done for convenience, though not necessary). One should note that we will treat the observed indicators as continuous, but the logic of what follows would hold if a link function were used. For example, $\boldsymbol{\nu}$ would be a vector of zeros when using a probit link, and the y_i's interpreted as y_i^*'s.

If one were to substitute the structural equation into the measurement equation, the result would be

$$\boldsymbol{y}_i = \boldsymbol{\lambda}(\gamma_1 t_i + \zeta_i) + \boldsymbol{\epsilon}_i \quad (7)$$

Given $E(\zeta_i) = 0$ and $E(\epsilon_i) = 0$,

$$E_i(\boldsymbol{y}_i) = E_i(\boldsymbol{\lambda}(\gamma t_i + \zeta_i) + \boldsymbol{\epsilon}_i) = \boldsymbol{\lambda}(\gamma_1 t_i) \quad (8)$$

Since $t_i = 0$ for the control group, the expectation of the observed outcome for the control group $E_i(\boldsymbol{y}_i)$ is simply an $N\ x\ 1$ vector of zeros. Meanwhile, $E_i(\boldsymbol{y}_i)$ for the treated group is

$$E_i(\boldsymbol{y}_i\ |t_i = 1) = \boldsymbol{\lambda}(\gamma(1)) = \boldsymbol{\lambda}\gamma \quad (9)$$

Given this result, one could solve for the true treatment effect, γ, by multiplying both sides of the equation by a $1\ x\ N$ vector where each element is simply one divided by the sum of the elements of $\boldsymbol{\lambda}$. For example, let's assume a three-item survey. If we limited only to those study participants in the treatment (recalling the mean true score of the control group is zero), we would have

$$E\begin{bmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} \gamma \quad (10)$$

Here, if the loadings are below one, the true treatment effect relative to the observed treatment effect will be larger. If the loadings are above one, the reverse is true. However, when we use a sum score model, we are constraining the loadings to be one, or at least to be the same across items (McNeish & Wolf, 2020). Therefore, with a misspecified sum score model, we would have a biased estimate of the treatment effect when the true loading is not equal to one. For example, if the loadings were all .7, the true treatment effect would be larger than the observed. But, if we were to constrain the loadings to be one as in a sum score model, the *estimate* of γ would equal the observed treatment effect and would be downwardly biased.

Impact of Observer Protocols

Up to now for simplicity we have focused on a RD that has a survey type measure as an outcome. Observer protocols introduce more complications because they introduce one or more rater(s) conducting the observation. This extra source of variation could exacerbate the issues of bias that we have outlined for the survey case. Sum scoring treats all items from each rater equally, a flawed assumption as demonstrated by Ho and Kane (2013), who showed that scores from observational protocols are influenced by various factors related to the rater and rating situation. In contrast, a model with multiple raters that takes into account differences between raters and items can distinguish between sources of variability, such as different contexts or rater perspectives, reducing the bias in the causal estimates.

While researchers in the social sciences have not commonly used outcomes assessed by multiple raters with RDs in the past, they often obtain ratings of a construct of interest provided by multiple raters and many largescale datasets include item responses from multiple raters. Researcher and evaluator motivations for involving multiple raters typically stem from two reasons connected to the varied rater scenarios in the introduction. The first is to gather different perspectives. For example, ratings of a child's position on a latent social-emotional learning continuum (e.g., self-efficacy) might be provided by teachers and parents (Bandura, 1994; Wheeler & Ladd, 1982) each which see a child in different contexts. The other is to improve the quality of information such as when the effectiveness of a given teacher's instruction might be evaluated by multiple observers, or classroom management practices by the students in the class (Schweig, 2014).

Research from the Measures of Effective Teaching (MET), a Gates Foundation-funded project emphasizes the importance of using multiple raters when observations are high stakes. A study using MET data conducted by Kane and Staiger (2012) revealed that roughly 37% of the

variability in raters' item responses was attributed to true trait-level differences in teacher effectiveness, while the rest was accounted for by factors such as the lesson, the rater, the section, and the interactions among those sources of construct-irrelevant variance. This means that in any single evaluation, a minority of the variance in an observational protocol score is the result of a target's behavior. Thus, using multiple ratings is essential when trying to detect causal effects.

However, the collection of multiple ratings brings new challenges. On the positive side, having multiple perspectives from a measurement tool can increase the reliability of the obtained scores by reducing the impact of individual evaluator biases. On the other hand, disagreement among raters, due to rater preferences and the context of ratings, is a frequent occurrence (De Los Reyes & Kazdin, 2005; Jones & Bergin, 2019; Styck et al., 2021), and researchers need models that can be used to estimate a true score in the face of such discrepancies in observed ratings, regardless of whether the discrepancies are due to rater error or contextual differences. Failure to do so has could lead to many potential measurement issues, including inaccurate Item Response Theory (IRT) person and item parameter estimates and inflated measurement reliabilities (Chen & Thissen, 1997; Jiao et al., 2005; Wilson & Hoskens, 2001). This leaves open the question of whether using multiple raters is more beneficial than using a single rater if an improper measurement model is used.

Proposed Model

Fortunately, many models have been developed to address this need and account for rater discrepancies. These include multi-trait multi-method (MTMM) models in an SEM framework (Geiser et al., 2019), as well as several IRT approaches, such as the many-facet Rasch model (Myford & Wolfe, 2003), the rater bundle model (Wilson & Hoskens, 2001), and the hierarchical

rater model (Patz et al., 2002). However, many of these measurement models are infeasible because they are misaligned with the purposes and situations that education policy researchers tend to use observational protocols. For example, the hierarchical rater model is designed for the case where there are several ratings from well-trained observers on a single instance with the goal of subtracting error that the raters might introduce while accounting for the dependency between the ratings (Patz et al., 2002). While this might be feasible in a clinical setting or in standardized test settings, where highly trained observers observe a limited number of subjects, it is much less viable in education evaluation and research where resources are constrained and observations are often made by people without extensive training on an observational protocol, like parent and teachers.

Accordingly, for this study we chose a model designed not only to account for multiple raters when producing scores, but also to be more conceptually understandable and easier to use when there are multiple raters who have different perspectives or are rating in different contexts (Bauer et al., 2013; Shin et al., 2019). This model has been developed independently under various names by several researchers, including the trifactor model by Bauer et al. (2013) and the third-order IRT model by Rijmen et al. (2014).

As presented by Bauer et al. (2013), the trifactor model decomposes the variance in item responses between two or more raters (e.g., a teacher and an independent researcher) into variance shared by both raters (a "common" factor), idiosyncratic to each rater (hereafter referred to as a "rater" factor), and shared by common items administered to both raters (hereafter referred to as an "item" factor). This easy-to-understand partitioning of variance makes it ideal for substantive education researchers who might hesitate to use more opaque models. The trifactor model can be estimated in either an SEM or IRT framework and aims to generate scores

that are free of rater bias, but also can be used to determine how much of the variance is due to rater or item differences.

We combine the trifactor model with the RD SEM model outlined by Soland et al. (2022) which is shown in Figure 1. This model is a fuzzy RD model that can integrate ratings from multiple raters as shown in Figure 2 below. In that figure, y_{1i1} through y_{6i1} are observed indicators of the construct of interest as rated by rater 1 while y_{1i2} through y_{6i2} are observed indicators of the construct of interest as rated by rater 2. Orthogonality constraints are imposed on all the factors in the model so that there is one common factor η_{Ci} that accounts for shared variance across all responses and is the outcome of interest, two rater factors (η_{1i} and η_{2i}) that account for the variance idiosyncratic to each individual rater, and 6 item factors ($S_{1i} - S_{6i}$) that account for the residual dependence due to the same item being completed by multiple raters. As in Figure 1, assignment to treatment z_i and related interaction terms represent the instrumental variable part of the fuzzy RD design and t_i represents treatment status. In Figure 2, unlike Figure 1, residual error terms ($\epsilon_1 - \epsilon_6$) are not shown. While Figure 2 shows a scenario with six shared items across two raters the model can account for scenarios with more than two raters or when all items are not shared. In those cases, only the items that are shared across raters have an item factor.

While this model combines other models proposed in the prior literature (trifactor [Bauer, 2013], SEM for RD designs [Soland et al., 2022]), little is known about how well it might perform under realistic conditions occurring in education research. In particular, the model's ability to recover true treatment effects has not been investigated. The following simulation studies are designed to examine the model's performance and, thereby, give researchers using multi-rater data in quasi-experimental settings a sense for whether it might be useful in their own

work. Secondarily, the simulations will demonstrate the utility of multiple raters in an RD design when using the RD Multi-rater model as well as popular simpler alternatives with single or multiple raters.

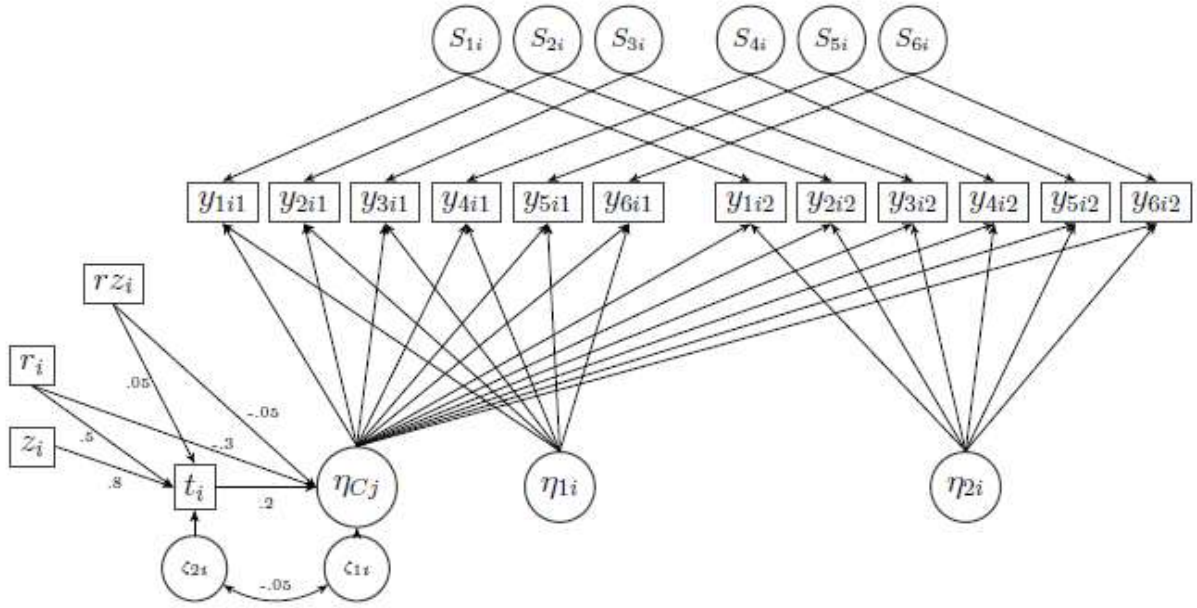

Figure 2. *Path diagram for the data-generating fuzzy RD model.*

Simulation Studies

In these two studies, we simulated RDs in an SEM framework and examined the potential improvements in the recovery of estimated treatment effects that come from incorporating information from multiple raters in a properly specified measurement model. The path diagram in Figure 2 served as our baseline generating model. As the figure shows (and according to the hypotheticals in the background section), there is a single latent variable of interest. That variable is measured using observed indicators, that are rated by two observers. There is a true

treatment effect of .2 (though we do vary the value of γ_i in the first simulation study). The loadings on the construct and observer latent variables as well as the residuals differ by item. While the figure presents a fuzzy RD (in which the correlation between the assignment variable and actual treatment status is .8), the same model was also used to create a sharp RD for the simulation studies, but with all paths from z_i and rz_i, the correlation between ζ_{1i} and ζ_{2i}, and the path from r_i to t_i constrained to zero. In each study below, we varied the values of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ or some combination of the two. The values of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ formed three conditions that were based on items of an observer protocol reported by Bauer et al. (2013), as well as six items from the CLASS (Pianta et al., 2008) and six items from the MQI (Hill et al., 2008) as reported by Kaine and Staiger (2012). These conditions are summarized in Table 1 below.

Table 1. Loading and Residual Variance Conditions

Condition 1 - Bauer et al. (2013) scale					
Item	Common λ	Rater 1 λ	Rater 2 λ	Item λ	Residual Var.
1	0.52	0.39			0.578
2	0.41	0.33			0.723
3	0.31	0.48			0.674
4	0.08	0.5			0.744
5	0.65	0.42			0.401
6	0.6	0.51	0.51	0.34	0.264
7	0.27	0.59	0.59	0.67	0.13
8	0.22	0.62	0.62	0.5	0.317
9	0.4	0.62	0.62	0.49	0.216
10	0.52		0.49		0.49
11	0.64		0.51		0.33
12	0.35		0.63		0.481
13	0.56		0.44		0.493
Condition 2 - CLASS from Kaine and Staiger (2012)					
Item	Common λ	Rater 1 λ	Rater 2 λ	Item λ	Residual Var.
1	0.75	0.18	0.18	0.2	0.365
2	0.51	0.28	0.28	0.69	0.185

3	0.1	0.35	0.35	0.58	0.531
4	0.9	0.1	0.1	0.14	0.16
5	0.2	0.29	0.29	0.7	0.386
6	0.4	0.4	0.4	0.1	0.67

Condition 3 - MQI from Kaine and Staiger (2012)					
Item	Common λ	Rater 1 λ	Rater 2 λ	Item λ	Residual Var.
1	0.6	0.2	0.2	0.2	0.56
2	0.51	0.5	0.5	0.1	0.48
3	0.05	0.35	0.35	0.51	0.615
4	0.4	0.1	0.1	0.35	0.708
5	0.2	0.29	0.29	0.6	0.516
6	0.3	0.601	0.601	0.7	0.059

These conditions reflect the two varied rater situations in which educational evaluators might use multiple raters with observational protocols identified in the introduction and background sections. Condition 1 represents a situation where perspectives from multiple raters, such as a parent and a teacher, are being collected. The scales used by each rater are slightly different and only four items are shared between them. These shared items are binary checklist-type items. The decision to include four shared items was made to align with the number of shared items used in previous studies that had similar scale lengths and multiple raters (Gresham & Elliott, 1990). In contrast, in Conditions 2 and 3, we simulate a scenario where multiple raters are

utilized in high-stakes evaluations, such as teacher evaluations, to ensure objectivity. In the high-stakes scenario, both raters use the same scale that includes all six items, which are more complex continuous items.

All conditions were replicated 1000 times in Mplus version 8.7 (Muthen & Muthen, 2017). Treatment effects based on simulated data were estimated using a Weighted Least Squares Means and Variance (WLSMV) adjusted estimator for the models that included categorical variables and a Maximum Likelihood (ML) estimator for those models with only continuous variables. As suggested by Bauer et al. (2013), the latent variables were scaled by standardizing the common factor and one of the rater factors, but freely estimating the mean and variance of the other rater factor.

Simulation Study 1. Feasibility Methods

In the first simulation study, we evaluated the capability of the RD Multi-rater model to recover different treatment effects with various sample sizes within the bandwidth. To do this, we generated simulated datasets with sample sizes ranging from 100 to 5000 participants within the bandwidth at intervals of 100, using the three loading and residual variance conditions specified in Table 1. These sample sizes covered a broad range of possible sample sizes, although the upper end would only be feasible with access to national or state level administrative data. In addition, we considered three different treatment effect sizes of 0, 0.1, and 0.2 SDs with a .2 SD effect size chosen to represent the upper limit of the majority of educational interventions that are well powered (Cheung & Slavin, 2016; Kraft, 2020; Rocconi & Gonyea, 2018). This range of true effect sizes allowed us to determine the potential for both false positives (Type I errors) and negatives (Type II errors/statistical power). After generating the simulated data, we then used the RD Multi-rater model to try to recover the true treatment effect.

Finally, model performance, including recovery of these estimated treatment effects, was examined in several ways. We first examined convergence rates, which were quantified by reporting the number of replications that converged out of the 1000 replications conducted per simulation. We then examined the mean treatment effect estimate (relative to the true data-generating parameters) and variance of the estimated treatment effects across all replications and sample sizes. Finally, we examined the Type I and II error rates of the treatment effect estimates.

Simulation Study 1. Feasibility Results

Convergence Rates.

Figure 3 shows convergence rates by sample size and loading condition for both sharp and fuzzy RDs. True treatment effect condition had no impact on convergence rates, so only convergence rates for the models with .2 treatment effect sizes are shown. As the figure shows, convergence rates vary between 667 and 998 out of 1000 replications when there are 100 observations within the bandwidth but improve steadily as the sample size increases. The extra parameters in the fuzzy design seemed to affect the ability to successfully complete all replications in two of the loading conditions even at higher sample sizes. This problem could likely be solved by using an alternate Bayesian approach.

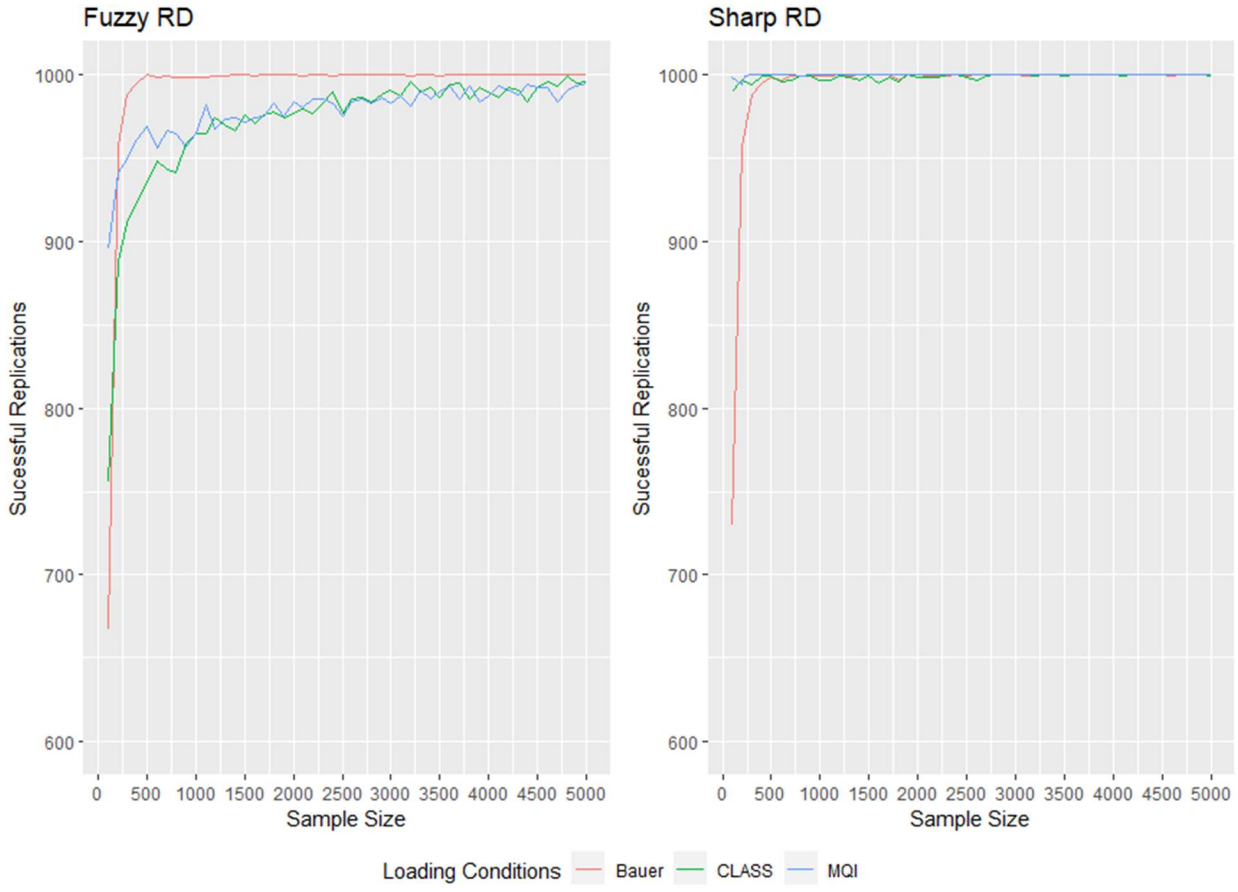

Figure 3. *Convergence Rate by Loading Condition (.2 Treatment Effect Size)*

Treatment Effect Recovery.

Figures 4-6 show the mean treatment effect estimate with bars representing 1 SD across replications for each loading condition for both sharp and fuzzy designs. For each model, all model parameters including the treatment effect were estimated using only the participants within the bandwidth. As seen below, bias was only a problem for smaller sample sizes and was most problematic in the fuzzy design. In particular, bias was present for the lower sample sizes of the Bauer condition in the fuzzy design where the bias started at 40% at sample size = 100.

However, by a sample size of 500, bias was negligible for every condition. After 500 participants, the variance of the treatment estimate was the main obstacle to the effective deployment of the RD Multi-rater model. For sharp RDs, around 800 participants in the bandwidth was sufficient such that there was no longer overlap between the error bars of all three treatment effects as seen in Figures 2-4, but it takes approximately 3000 more participants so that the error ribbons no longer overlap for a treatment effect of .1 and .2[4]. For fuzzy RDs it took approximately 1300 participants so that there is no longer overlap between a treatment effect of 0 and .2, and even with 5000 participants, there was still overlap between the SDs of the estimated treatment effect of .1 from .2 in the Bauer condition (Figure 2). The greater variance in the fuzzy design likely occurred because the use of an IV introduces another source of variability. If the correlation between *z* and *t* was stronger than the .8 value used in our simulation the differences in the variance between the Fuzzy RD and Sharp RD would likely be smaller and conversely a weaker correlation would have magnified the differences.

[4] The overlapping of CIs does not necessarily imply that the difference between two statistics is statistically significant (Knezevic, 2008). In this case it is being used to show at what sample sizes you may get the same point estimate for different true effect sizes.

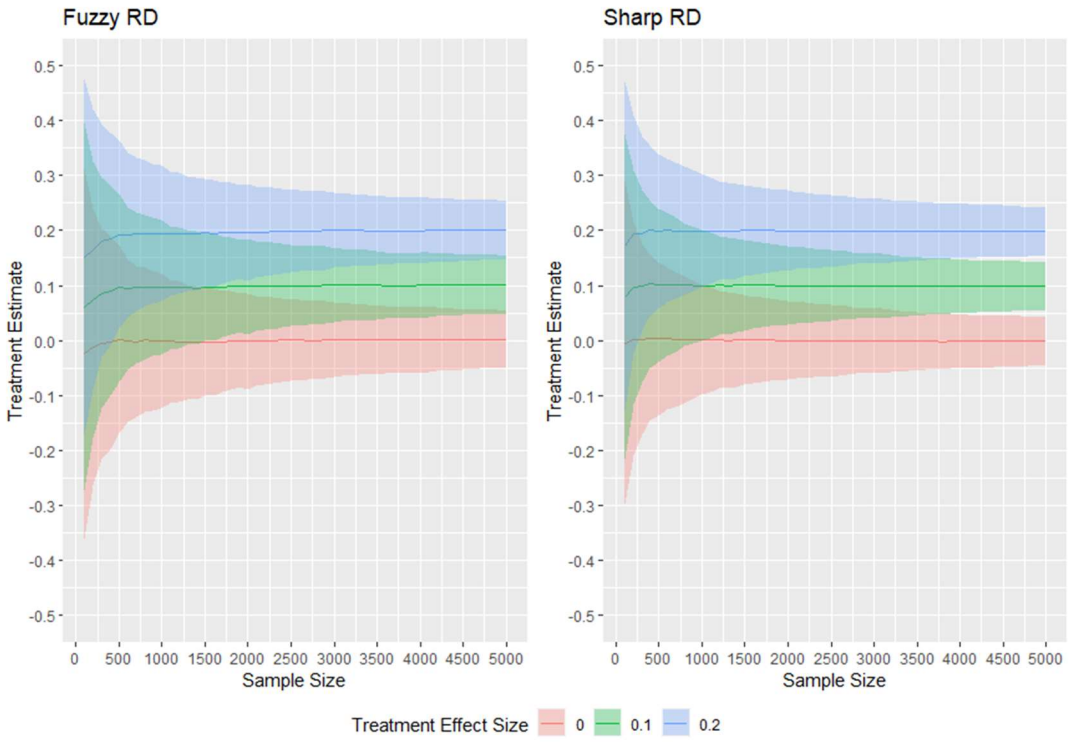

Figure 4. *Mean Treatment Estimate with 1 SD Error Bar by Sample Size (Bauer)*

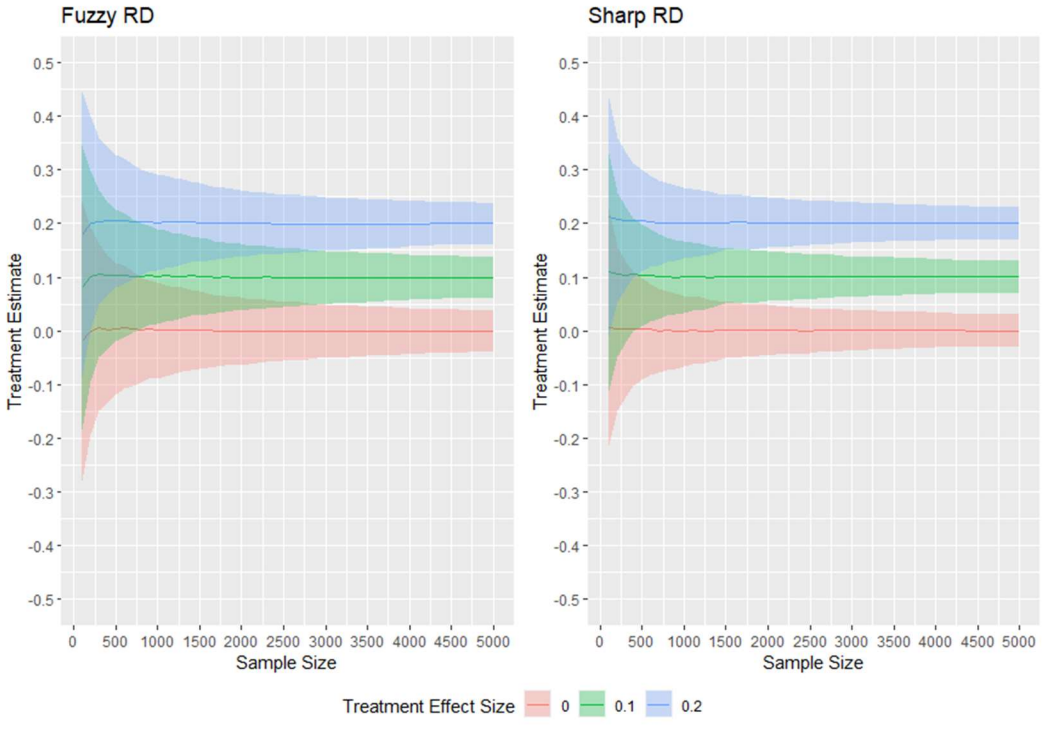

Figure 5. *Mean Treatment Estimate with 1 SD Error Bar by Sample Size (CLASS)*

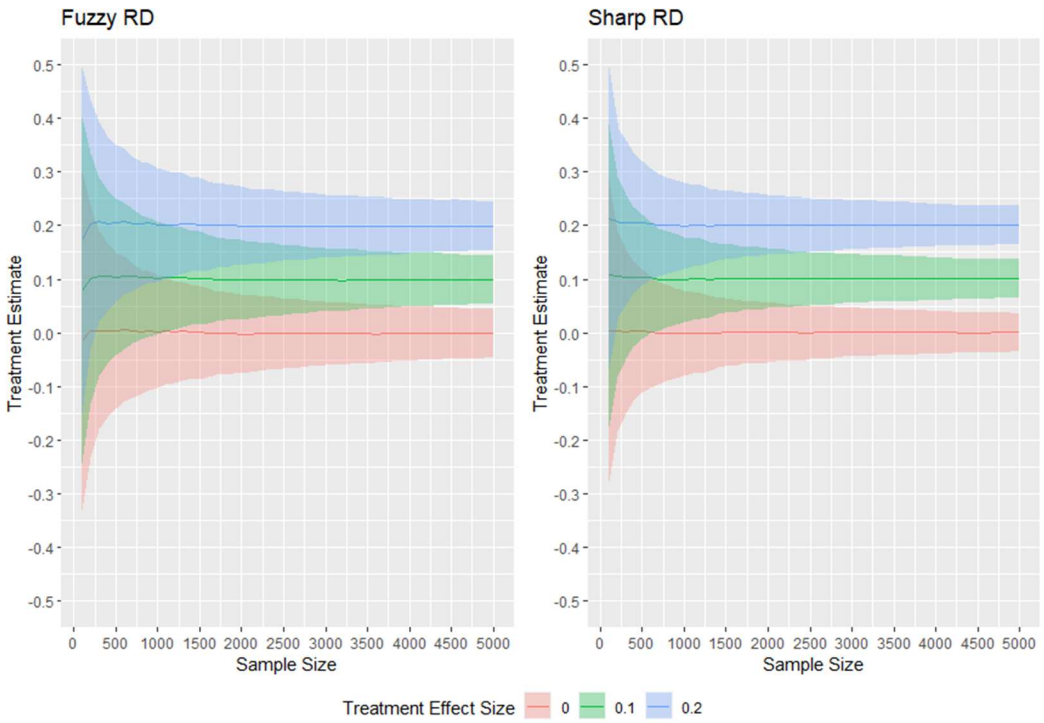

Figure 6. *Mean Treatment Estimate with 1 SD Error Bar by Sample Size (MQI)*

Type I and II Errors.

To further investigate how variability in the estimates affected results, the proportion of significant results ($p<.05$) are shown in Table 2. There were relatively low Type I error rates (< 1.2%) that tended to be greater for Fuzzy RDs, but varied over sample size and loading condition. In contrast, the Type II error rates were higher and clearly driven by true effect size, sample size, loading condition, and type of research design (fuzzy versus sharp). Larger effect sizes and sample sizes had lower Type II error rates while Sharp RDs had lower Type II error rates than Fuzzy RDs. For loading conditions, the CLASS condition had the lowest Type II error rate likely because both raters rated all items, and its loadings resulted in lower levels of residual variance. The condition with the next lowest Type II error rate, the MQI condition, also had all

items rated by both raters but explained less variance. Finally, the Bauer condition had the highest Type II error rate, likely because only 4/13 items were rated by both raters. Thus, while sample sizes and effect sizes were important the exact observational protocol used also drove Type II error rates.

Table 2. Percent Significant Treatment Effect Coefficients for Selected Sample Sizes

N	Effect Size	Prop. Significant Fuzzy RD			Prop. Significant Sharp RD		
		Bauer	CLASS	MQI	Bauer	CLASS	MQI
500	0	0.062	0.044	0.053	0.057	0.046	0.059
500	0.1	0.113	0.145	0.126	0.117	0.2	0.15
500	0.2	0.238	0.39	0.291	0.307	0.577	0.459
1000	0	0.058	0.061	0.062	0.062	0.047	0.045
1000	0.1	0.143	0.218	0.177	0.184	0.311	0.231
1000	0.2	0.386	0.64	0.495	0.536	0.847	0.704
2000	0	0.064	0.067	0.05	0.071	0.053	0.049
2000	0.1	0.197	0.368	0.281	0.315	0.59	0.434
2000	0.2	0.641	0.892	0.785	0.81	0.987	0.945
5000	0	0.047	0.06	0.062	0.055	0.05	0.05
5000	0.1	0.473	0.732	0.589	0.619	0.927	0.807
5000	0.2	0.956	1	0.99	0.991	1	1

Overall, the first simulation study showed that the RD Multi-rater model can provide unbiased estimates on average, but that it might take relatively large sample sizes within the bandwidth to provide unbiased estimates. The requirement for a large sample size may be mitigated using cases outside the bandwidth to estimate some parameters (and, in particular, measurement model parameters [Soland, Johnson, & Talbert, 2022]), but a more in-depth examination of such considerations is beyond the scope of this study. While raw sample size is important, other aspects like how well the items load on to the latent construct, how many items are shared, and the research design itself can drive the ability to detect treatment effects. These factors drove reductions in Type II error rates that ranged from .4% to 25.1%, and sometimes exceeded the reductions in Type II error rates that accompanied doubling the sample size (3% to 27.4%).

Simulation Study 2. Comparative Benefit Methods

In the second study, we compared the RD Multi-rater model with misspecified models that included both raters or discarded one, on recovery of treatment effect parameters. Two types of misspecified models were examined: a model in which a sum score model is wrongly fit, and a measurement model that treats raters as exchangeable. Figure 2 was the basis for the data generation with the true treatment effect set to .2 SDs. Once again, all three conditions in Table 1 were used and loadings were changed accordingly. As in the first simulation study, treatment effects were first estimated using a SEM model that integrated multiple ratings, the RD Multi-rater model. We then fit a similar model, but used a measurement model that mimicked sum scores to represent the dependent variable. To avoid scale indeterminacy in the dependent variable that can result from using sum scores, the sum scores were produced by fitting a highly constrained measurement model akin to the one described by McNeish and Wolf (2020). Specifically, we constrained all the loadings equal and set ($\theta_{11} = \theta_{22} = \theta_{33} = \theta_{44} = \theta_{55} =$

θ_{66}). To ensure this approach matched the use of actual sum scores, we produced scaled mean scores based on the generated item responses, and verified that they corresponded to scaled factor scores produced using our highly constrained SEM (that is, we made sure they had a correlation equal to one). Finally, to examine the case where rater differences are ignored, we estimated treatment effects using the simple latent variable model in Figure 1. That model effectively ignores the distinction between the two raters while still fitting a measurement model that puts latent variables on a scale comparable to the ones used in the other conditions (Horton & Fitzmaurice, 2004; Kraemer et al., 2003; van Dulmen & Egeland, 2011).

All in all, there were three models were used to analyze the datasets and two of the models had variants where one or both raters were used. This meant there were five analysis conditions: the sum score all raters (SS-all), sum score one rater (SS-1r), simple latent variable all raters (SLV-all), simple latent variable one rater (SLV-1r), and the RD Multi-rater (RDMR). These conditions are summarized below in Table 3. Given results from study 1, the sample size within the desired bandwidth was set equal to 1,000. We evaluated the accuracy of each model's estimated treatment effects by comparing their means, variances, and Type II error rates.

Table 3. Analysis Conditions used in Simulation Study 2

Condition	Abbreviation	Description
Sum Score (All raters)	SS-all	This condition takes the ratings from each rater and pools them together by weighting each item regardless of rater equally.

Sum Score (One rater)	SS-1r	This condition discards the ratings from the second rater and weights equally the items scored by the first rater.
Simple Latent Variable	SLV-all	This condition takes the ratings from each rater and pools them together by estimating the loading of each item separately and then using them to estimate a latent variable.
Simple Latent Variable (One Rater)	SLV-1r	This condition discards the ratings from the second rater then estimates the item loadings for the items scored by the first rater to estimate a latent variable.
RD Multi-rater	RDMR	This condition treats the ratings from each rater as distinct and uses the difference between the two sets of ratings to partition the variance between the target latent variable, item latent variables, and rater latent variables.

Simulation Study 2. Comparative Benefit Results

Treatment Effect Mean and Variance Results.

Figure 7 shows a plot with the average estimated treatment effect across all 1,000 replications with 1SD error bars by loading condition and scoring method for a sample size of 1000. Several

aspects of the figure are worth noting. For one, the variability of the results tends to be smaller for the sum score results for the items that use continuous measures than any of the SEM models. This smaller variance likely occurs because (a) the sum score constrains all the loadings to one, which would reduce the estimated variance of the latent variable because each item is always weighted the same (b) there is error in the loadings when they are estimated rather than fixed, and (c) the model assumes any common score is identical regardless of the items that produced it. Both the SLV and sum score models had much smaller variances than the RD Multi-rater model, likely because the number of parameters estimated in the RD Multi-rater model was much larger.

Another takeaway is that we see a large downward bias in the treatment effect estimates when using sum scores and the simple latent variable model due to model misspecification. For the sum score models, there are two misspecifications: (1) the item loadings are fixed for each item to one when they are below one and not equal and (2) the multiple rater structure is ignored. For the simple latent variable model, only the multiple rater structure was ignored, but there were still large amounts of bias that ranged from .045 to .13 SDs out of .2 SDs. The SLV models generally had less bias than the sum score models, but the results were inconsistent. In the Bauer condition, sum score models showed less bias than SLV models. This was likely because ignoring the rater structure resulted in item loadings that were farther from those in the true measurement model compared to the fixed loadings of a sum score. The RD Multi-rater model, on the other hand, showed minimal bias (<.005 SDs) across all three loading conditions in both fuzzy and sharp RDs.

A third takeaway is that using two raters with a misspecified measurement model decreased the variance of the estimate, but did not consistently reduce the bias of the treatment estimate. For

most of the loading conditions the treatment estimates were very similar between the models that used all and one rater for both the sum score and SLV models likely because the raters had similar means, variances, and loadings. In contrast, the RDMR model which used the information from the two raters more effectively had minimal bias. One of the primary theoretical benefits of using multiple raters is that it provides more information that can lead to better and more accurate estimates. Our results indicate that using a missspecified measurement model can negate these benefits.

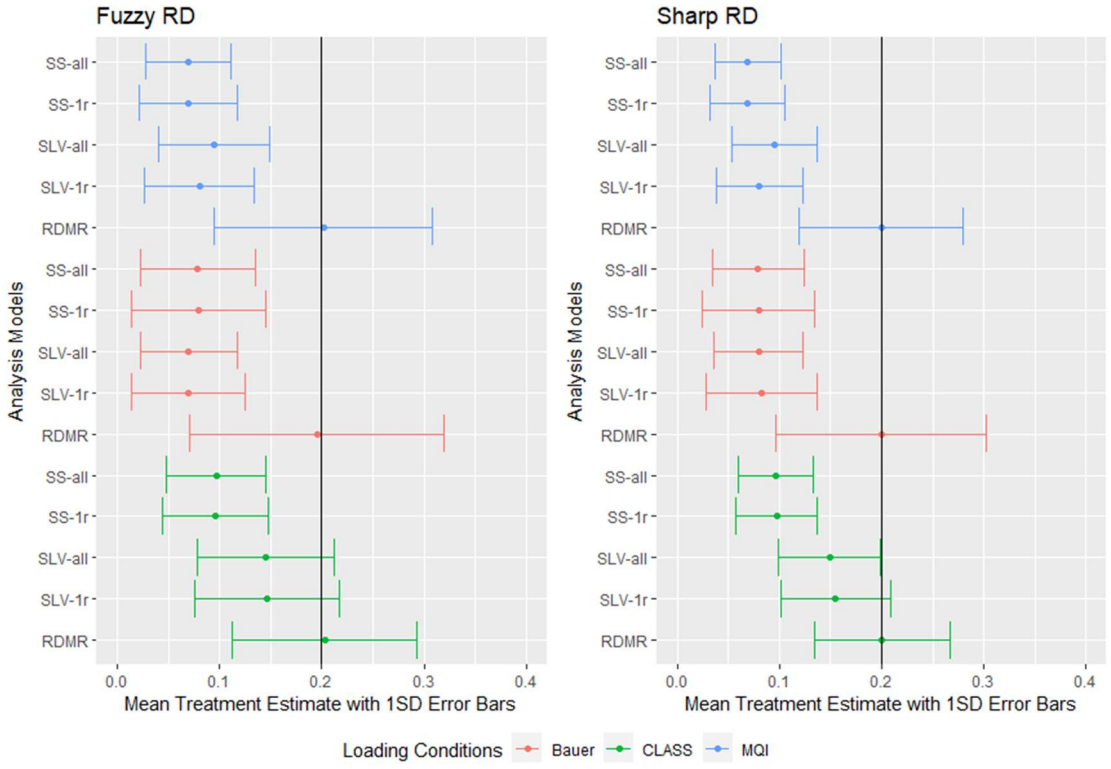

Figure 7. *Mean Treatment Estimate with 1 SD Error Bars by Analysis Model*

Type II Error Results.

Both the variance and bias influenced Type II error rates for each analysis model, which are shown in Figure 8. Despite the variability of the RD Multi-rater model, it had the least Type II errors of the analysis models. This was the result of the large amounts of bias in the other

analysis models, which outweighed the accompanying reduction in variability. Only in the Sharp RD, CLASS condition did any other model have equivalent Type II error rates (SLV-all). On average, in the fuzzy designs, the RD Multi-rater model resulted in 9% less Type II error rates than the all-rater sum score model and 3% less Type II error rates than the SLV-all model. In sharp designs there was a marginally better improvement with the RD Multi-rater model producing 10% less Type II error rates than the SS-all model and 4% less Type II error rates than the SLV-all model.

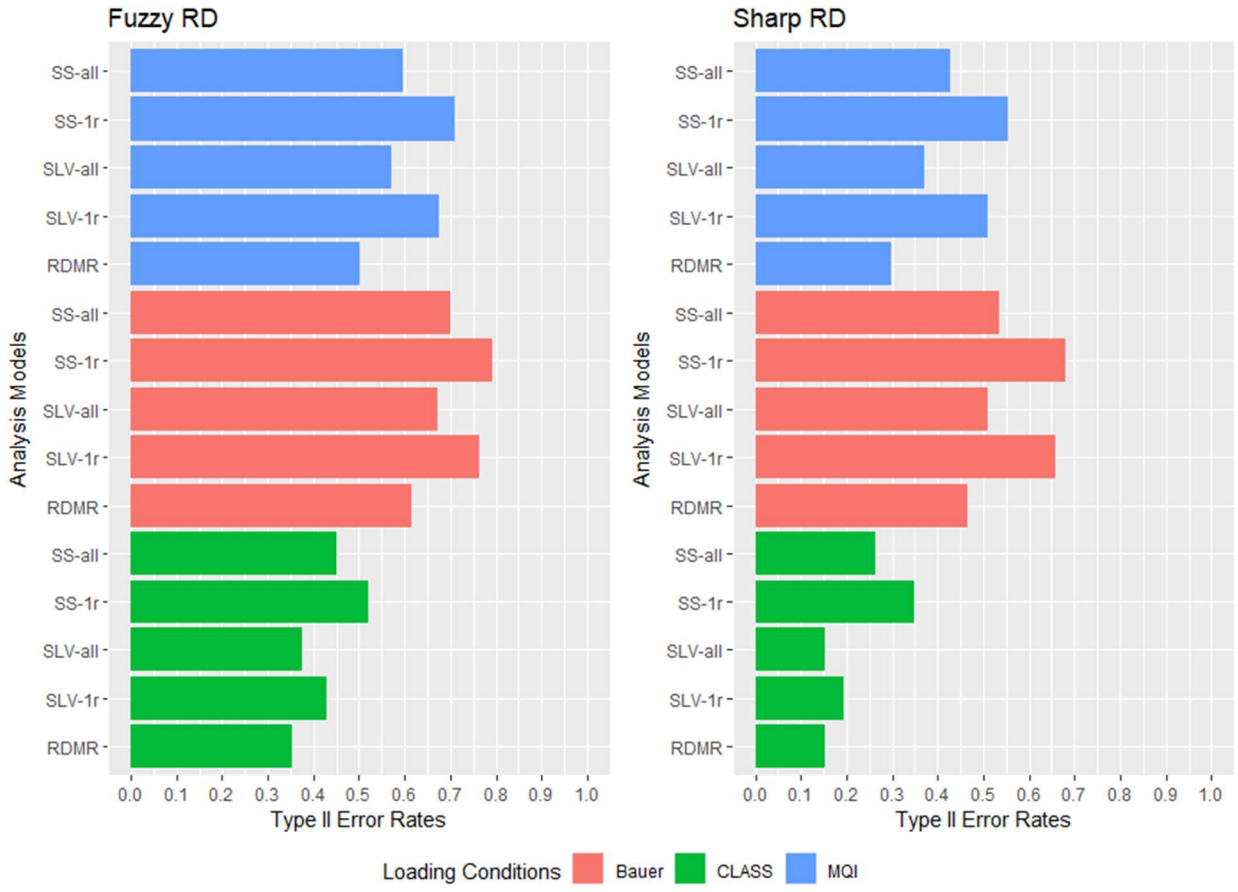

Figure 8. *Type II Error Rates by Analysis Model*

Discussion

The use of RD designs in education is growing in popularity, and they have been applied to observer protocol data in the past. However, when working with multiple raters in observer protocols, challenges arise in terms of how to effectively combine ratings and accurately estimate the target latent variable. The simple approach of using sum scores, relies on a number of oftentimes unjustifiable assumptions (McNeish & Wolf, 2020) and has been shown to bias more simple multi-item measures like survey scales (Soland et al., 2022). When dealing with more complex data from multiple raters in observational protocols, it is likely that the same or greater biases will occur when using the sum score approach. In our study, we quantify these biases while investigating the feasibility of a possible solution, the RD Multi-rater design. Results provide several insights that are relevant to applied researchers and program evaluators.

First, we determined that the RD Multi-rater design requires large samples to be adequately powered to detect medium and small effects. These sample sizes may be difficult to obtain. Given the impact of loadings, shared items, and design on the variance of the estimated treatment effect the RD Multi-rater model is most effective with a sharp design and with raters using the same well validated scale. Ideal sample sizes for a fuzzy RD would be greater than 3000 but the model can provide valuable insights especially with well-designed observer protocols with a sample size of 500. In a brief follow up analysis using the CLASS condition with a fuzzy RD we showed the large sample size requirement was not merely the result of more parameters being estimated. Even when measurement model parameters were fixed there was minimal impact on the treatment effect standard error. This may indicate that the large variance-bias tradeoff comes from properly modeling observer protocols, which are noisy measures of a target construct.

However, we show that even when not optimally powered using a sample size of a 1000 the RD Multi-rater model is superior to four other approaches of using multi-rater data. We demonstrated that using a sum score model when its assumptions are violated can induce substantial bias into treatment effect estimates. Further, under the conditions we used, those treatment effect estimates were downwardly biased. In both sharp and fuzzy RD designs, wrongly fitting a sum score model resulted in understating a true treatment effect of .2 units by .1 to .13 units depending on the true loadings in the model. In other words, using a sum score model when its assumptions conflicted with a complex multi-rater structure led to a downward bias of over 50% of the true treatment effect.

Similar problems were present when using a simple latent variable model that ignored rater structure. While the additional flexibility of allowing item loadings to vary by item meant that there was a bias of only .05 units in the CLASS condition the bias reached .13 units in the Bauer condition. The large biases reflected the major distortion that ignoring the multiple rater structure inflicts. Our results indicate that simply treating observer protocols as if they were survey question is likely to be inadvisable.

Unsurprisingly, the large amounts of bias also increases Type II error rates. Thus, even if one does not care about the point estimate of a treatment effect, only whether it is significant or not, using an improper measurement model with can be hugely problematic. By contrast, estimating the RD treatment effect with the RD Multi-rater model led to practically no bias, on average, and much lower Type II error rates.

Furthermore, our examination of bias showed that the models that include one rater produced mostly equivalent treatment estimates on average to their counterparts that include both raters. Practically this means that the additional ratings, which are often resource intensive to obtain, did

not provide a meaningful boost in accuracy and only decreased variability slightly. Thus, using multiple ratings or conducting multiple ratings in the RD context provide very little benefits without using a properly specified measurement model. Averaging the two ratings may provide a false impression that the resulting product will be more accurate, and although this is a finding specific to the RD design it draws into question the use of observer protocols in high stakes' evaluations.

All told, these results provide a strong initial argument in favor of attempting to estimate RDs using the RD Multi-rater model when multi-rater data is available. Even when sample sizes are small the model still successfully ran in most replications and the model can provide additional information about the likely range of treatment effects. Furthermore, we revealed that using other simpler models can drastically underestimate the treatment effect and that using an overly simplified measurement model negates the benefit of using multiple raters. Given the amount of time that researchers and program evaluators invest in power calculations prior to evaluations, paying equal attention to measurement issues appears warranted. In fact, our results indicate that poor measurement could, in many cases, wipe out the benefits of substantially increasing sample sizes in RDs and the considerable expense of using multiple raters.

While this is a strong initial argument for the use of the RD Multi-rater model, a major obstacle is the large sample sizes required to reduce variance. For RD designs, a sample size with over 1000 observations in the bandwidth might not be feasible. Fortunately, there are plausible improvements that can be implemented and could reduce the sample sizes needed greatly. In particular, observations outside the bandwidth could be used to calibrate the parameters for the tri-factor portion of the RD-multi rater model. This would free that portion of the model from the

constraints of only using bandwidth observations, while keeping the benefits of using a bandwidth for estimating the treatment effect.

Limitations and Future Directions

Our study has a few limitations that bear mention. First, the simulation conditions covered are not exhaustive and do not consider other potential measurement problems that might occur. For example, other research on the tri-factor model has shown the impact of raters with drastically different loadings (Soland & Kuhfeld, 2022), which was not covered in this study. Second, the dearth of easily accessed empirical data that includes multiple raters in RD design means that we were unable to show how using the RD Multi-rater model might affect a study's conclusions. Thus, the findings from our simulation studies should be interrogated using empirical data.

Going forward, as mentioned above, our results also suggest that better methods need to be developed for observational protocols and RD designs that work for smaller sample sizes. Besides using observations outside the bandwidth to calibrate parts of the model, research around methods that employ Bayesian or other approaches that can utilize information from other datasets should be explored. Future research could also evaluate whether other measurement models are more appropriate for select applications than the tri-factor model we used for this study.

Finally, this study assumes that the true data generating model for observation protocols is adequately represented by the tri-factor model. While we outlined why we believe that the tri-factor model is appropriate, it is impossible to determine the true measurement model of empirical data. This means that the gains in accuracy in this study are a ceiling. If the tri-factor model is a significant misspecification of the data generating model for the empirical data it is utilized with, there is a likelihood of bias. However, the alternative approaches we investigated

are unlikely to be better specified, which suggests that the comparative benefits observed in our study will persist.

Conclusion

Our study highlighted the challenges of using observer protocol data from multiple raters in education research, specifically in the context of regression discontinuity (RD) designs. The traditional approach of using sum scores is shown to be biased and inaccurate, and an RD Multi-rater model is proposed as a possible solution. The study finds that the RD Multi-rater model requires large sample sizes to be adequately powered but is still superior to other approaches when not optimally powered because improper measurement models lead to substantial bias and increased Type II error rates. Additionally, the study suggests that simply averaging multiple ratings may provide a false impression of accuracy and that poor measurement could wipe out the benefits of increasing sample sizes in RD designs. Finally, the study suggests improvements to the RD Multi-rater model that could reduce the required sample sizes.

Chapter 3 References

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.

Bandura, A. (1994). Self-efficacy: Wiley Online Library. *Google Scholar*.

Bassok, D., Dee, T. S., & Latham, S. (2019). The effects of accountability incentives in early childhood education. *Journal of Policy Analysis and Management*, *38*(4), 838–866.

Bauer, D., & Curran, P. (2015). The discrepancy between measurement and modeling in longitudinal data analysis. *Advances in Multilevel Modeling for Educational Research: Addressing Practical Issues Found in Real-World Applications*, 3–38.

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475.

Berk, R., 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*(2), 191–208.

Black, D., Galdo, J., & Smith, J. A. (2005). Evaluating the regression discontinuity design using experimental data. *Unpublished Manuscript*.

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

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265–289.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Danielson, C. (2014). Framework for teaching. *Adapted for the Kentucky Department Of*.

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin*, *131*(4), 483.

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

Deke, J., & Dragoset, L. (2012). Statistical Power for Regression Discontinuity Designs in Education: Empirical Estimates of Design Effects Relative to Randomized Controlled Trials. Working Paper. *Mathematica Policy Research, Inc.*

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*, 241–261.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Geiser, C., Hintz, F. A., Burns, G. L., & Servera, M. (2019). Structural equation modeling of multiple-indicator multimethod-multioccasion data: A primer. *Personality and Individual Differences*, *136*, 79–89.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37*(3), 447–456.

Gorter, R., Fox, J.-P., Apeldoorn, A., & Twisk, J. (2016). Measurement model choice influenced randomized controlled trial results. *Journal of Clinical Epidemiology*, *79*, 140–149.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system: Manual*. American guidance service.

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.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*(4), 430–511.

Ho, A. D., & Kane, T. J. (2013). The Reliability of Classroom Observations by School Personnel. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

Horton, N. J., & Fitzmaurice, G. M. (2004). Regression analysis of multiple source and multiple informant data from complex survey samples. *Statistics in Medicine*, *23*(18), 2911–2933.

Hutcheon, J. A., Harper, S., Liauw, J., Skoll, M. A., Srour, M., & Strumpf, E. C. (2020). Antenatal corticosteroid administration and early school age child development: A regression discontinuity study in British Columbia, Canada. *PLOS Medicine*, *17*(12), e1003435. https://doi.org/10.1371/journal.pmed.1003435

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*.

Jones, E., & Bergin, C. (2019). Evaluating teacher effectiveness using classroom observations: A Rasch analysis of the rater effects of principals. *Educational Assessment*, *24*(2), 91–118.

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

Knezevic, A. (2008). Overlapping confidence intervals and statistical significance. *StatNews: Cornell University Statistical Consulting Unit*, *73*(1).

Kolesár, M., & Rothe, C. (2018). Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, *108*(8), 2277–2304.

Kraemer, H. C., Measelle, J. R., Ablow, J. C., Essex, M. J., Boyce, W. T., & Kupfer, D. J. (2003). A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry*, *160*(9), 1566–1577.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods*, *7*(2), 228.

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.

Marsh, H. W. (1989). Confirmatory factor analyses of multitrait-multimethod data: Many problems and a few solutions. *Applied Psychological Measurement*, *13*(4), 335–361.

McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, 1–19.

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

Muthen, L. K., & Muthen, B. (2017). *Mplus Version 8 User's Guide*. Muthen & Muthen.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Network, N. E. C. C. R. (2006). *NICHD Study of Early Child Care and Youth Development (SECCYD)*. National Institutes of Health, National Institute of Child Health and Human ….

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System™: Manual K-3.* Paul H Brookes Publishing.

Pohl, S., & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*, *45*(1), 45–72.

Rijmen, F., Jeon, M., Von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, *39*(4), 235–256.

Rocconi, L. M., & Gonyea, R. M. (2018). Contextualizing Effect Sizes in the National Survey of Student Engagement: An Empirical Analysis. *Research & Practice in Assessment*, *13*, 22–38.

Schweig, J. (2014). Cross-level measurement invariance in school and classroom environment surveys: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, *36*(3), 259–280.

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.

Shin, H. J., Rabe-Hesketh, S., & Wilson, M. (2019). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*, *54*(3), 360–381.

Soland, J., Johnson, A., & Talbert, E. (2022). Regression Discontinuity Designs in a Latent Variable Framework. *Psychological Methods*.

Soland, J., & Kuhfeld, M. (2022). Examining the performance of the trifactor model for multiple raters. *Applied Psychological Measurement*, *46*(1), 53–67.

Styck, K. M., Anthony, C. J., Sandilos, L. E., & DiPerna, J. C. (2021). Examining rater effects on the classroom assessment scoring System. *Child Development*, *92*(3), 976–993.

van Dulmen, M. H., & Egeland, B. (2011). Analyzing multiple informant data on child and adolescent behavior problems: Predictive validity and comparison of aggregation procedures. *International Journal of Behavioral Development*, *35*(1), 84–92.

Wheeler, V. A., & Ladd, G. W. (1982). Assessment of children's self-efficacy for social interactions with peers. *Developmental Psychology*, *18*(6), 795.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics*, *26*(3), 283–306.