Investigating Sources of Treatment Effect Heterogeneity in Intervention Research

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Education

have examined a dissertation entitled

Investigating Sources of Treatment Effect Heterogeneity in Intervention Research

presented by Catherine Armstrong Asher

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. James Kim

Signature _____

Typed name: Prof. Luke Miratrix

Signature _____

Typed name: Prof. Martin West

Date: May 10, 2021

Investigating Sources of Treatment Effect Heterogeneity in Intervention Research

A Dissertation

presented by

Catherine Armstrong Asher

to

The Committee on Higher Degrees in Education

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy in

the subject of

Education

Harvard University

Cambridge, Massachusetts

May 2021

Dissertation Advisor: Professor James S. Kim Catherine Armstrong Asher

Investigating Sources of Treatment Effect Heterogeneity in Intervention Research

Abstract

Intervention research in education investigates whether and how particular educational programs help students learn. Treatment effect heterogeneity, or variation in a program's impact, makes this effort more complicated, because it means more work must be done to understand what works, for whom, and in what contexts. This dissertation consists of three papers that explore distinct sources of treatment effect heterogeneity in the context of educational interventions.

The first paper, co-authored with Ethan Scherer, uses a factorial design to compare the effectiveness of three features in a reading-focused text messaging intervention: personalized information about students' observed reading behaviors, goal setting for summer reading, and framing the purpose of reading as useful for skill-building (an instrumental view), a fun activity (an entertainment view), or both. We find that that personalized messages increase the amount of summer reading and students' reading skills when they return to school in the fall. The effects of personalized information on test scores are amplified when families receive a combination of instrumental and entertainment-framed messages. We find no evidence of impacts for goal setting.

The second paper investigates how the effects of a district-run, universal-access pre-K program vary based on the access of the control group to different counterfactual options. Using administrative records and program waitlists, I reconstruct a matched sample that was plausibly randomized, either to receive a space in the district pre-K program or not. While on average I

find no evidence of an impact of the pre-K program on kindergarten readiness, I find large positive effects in communities where the control group students are more likely to report either not attending pre-K or attending a different subsidized pre-K program.

The third paper considers an emerging experimental design in education research: Sequential Multiple Assignment Randomized Trials (or SMARTs), which are used to develop, refine, and test adaptive interventions. I use principal stratification to identify endogenous subgroups determined by how individuals respond to each of the Phase 1 treatments, and I use these strata to highlight a latent assumption in current analytic techniques. I also present a simulation study to explore the magnitude of estimation bias when this assumption is violated.

Table of Contents

Title Page……..i
Copyright……ii
Abstract……...iii
List of Figures……vi
List of Tables……vii
Acknowledgments……viii
Dedication……...xi
Introduction……..1

Paper 1: Using a Factorial Design to Maximize the Effectiveness of a Parental Text Messaging Intervention……..6
- Background……..6
- Intervention……11
- Methods……..15
- Results……22
- Mechanisms and Student-Level Moderators……30
- Discussion……..34

Paper 2: Investigating How District-Provided Pre-K Impacts Depend on the Counterfactual…..43
- Background……43
- Methods……..49
- Results……57
- Discussion……..69

Paper 3: Inference in Sequential Multiple Assignment Randomized Trials using Principal Stratification……...74
- Background……74
- Modeling Responsiveness Using Potential Outcomes……...81
- Simulations……87
- Discussion……..96

Conclusion……..100
Appendices: Additional Materials……..……103
References……107

List of Figures

Figure 1.1 Predicted Reading Behaviors, by Personalization and View-of-Reading Conditions 24
Figure 1.2 Predicted Test Scores, by Personalization and View-of-Reading Conditions…….29

Figure 2.1 Demographics of District Pre-K Attendees Over Time…..……..58
Figure 2.2 Control Group Experiences with Pre-K Over Time…….59
Figure 2.3 Average Kindergarten Readiness Over Time, by School Community…….62
Figure 2.4 Average Kindergarten Readiness Among the Control Group, by School Community 69

Figure 3.1 Example SMART Design with Re-Randomization for Non-Responders……77
Figure 3.2 Bias in Estimating Phase 1 ATE……..94
Figure 3.3 Bias in Estimating Phase 2 ATE……..94
Figure 3.4 Bias in Estimating Phase 1 ATE with Heterogeneous Phase 2 Effects……95
Figure 3.5 Bias in Estimating Phase 2 ATE with Heterogeneous Phase 2 Effects……96

Figure B.1 Flow Chart Showing Construction of Analytic Samples, Paper 2…….105

List of Tables

Table 1.1 Baseline Characteristics and Balance Checks……...16
Table 1.2 Treatment Conditions in Factorial Design with Effect Coding…….18
Table 1.3 Differential Effects of Text Messaging Components on Parent Outcomes……...23
Table 1.4 Differential Effects of Text Messaging Components on Student Test Scores..…….27
Table 1.5 Differential Effects of Text Messaging Components on Alternate Student Outcomes 31

Table 2.1 Demographic Characteristics of Pre-K Attendees and Students Who Joined District in Kindergarten……..51
Table 2.2 Demographic Characteristics and Balance Checks from Quasiexperimental Analysis 55
Table 2.3 Pre-K Experience by School Community Among All Kindergarten Students……..60
Table 2.4 Effect of Attending District Pre-K on Kindergarten Readiness……64
Table 2.5 Impact Estimates by School Community……...66
Table 2.6 Balance Checks from Lottery-Based Sample, by School Community……..68

Table 3.1 Principal Strata and Potential Responsiveness in a SMART…….82
Table 3.2 Potential Outcomes under Example SMART Design with Re-Assignment for Non-Responders…….83
Table 3.3 Principal Strata and Responder Indicators in Simulations……88
Table 3.4 Parameterization of Simulated Potential Outcomes……..89
Table 3.5 Simulation Scenarios…….90

Table A.1 Text Message Topics Organized by Theme……103
Table A.2 Comparing Message Variations Across Conditions for Two Example Messages….104
Table B.1 Demographic Characteristics for Students with Known and Unknown Pre-K Experiences……..106

Acknowledgments

I am grateful for the support of the many individuals and groups who helped make this dissertation possible. First, I am grateful for the funding provided by the Institute for Education Services training grant R305B150012 provided to Harvard University and administered through the PIER program. Additionally, I am appreciative of my research partners, particularly the Georgia school district in Paper 2, whose staff were so helpful in providing information about the district's pre-K program and data systems and providing permission for me to conduct this research study. I have deeply enjoyed collaborating with Ethan Scherer on Paper 1 – Ethan, the paper is stronger, and every phase of the research process was more fun because we were working together.

I would also like to acknowledge the community of support at HGSE and beyond that helped me throughout my doctoral journey. Our PhD cohort has come so far since orientation. You all have defined my time at HGSE, and I am so grateful for the ways that I have learned from you over the past six years and for the ways we have always prioritized lifting each other up. I am especially grateful to my Prosem and Comps Reading Groups for the camaraderie and shared workload, and for our EPPE community that has been in the trenches through the coursework, the milestones, and the general existential doubts: Kirsten, Masha, Ben, Emily, Sophie. Monnica and Eddie – thank you for both holding me accountable and making me more productive on this dissertation journey, but also giving me love and understanding when I struggled. This pandemic has been so much less lonely, and this dissertation has actually happened, because of you two and our writing group.

The READS Lab and CARES Lab have brought so much time my time at HGSE. Thank you all for giving me spaces think in new ways and opportunities receive feedback, and for being

friendly and caring souls! Thank you to Alexa Seney and Laura Peters for your Zumba and TBC classes – your positive energy and excellent music selections have provided weekly bright spots over the past six years. Through the struggles of my first-year coursework to the ongoing Covid pandemic, you have both been integral to my physical and mental health in grad school.

I am grateful to my family, particularly my parents and my sister, who have tolerated my multiple-times-a-week phone calls for 6 years and celebrated every step of progress I made during this program. And, of course, I am so grateful for the countless contributions of my husband. Ben, you have been the steadiest, most consistently supportive partner for the past six years. You never doubted me (even when I doubted myself), you celebrated my successes, helped me get through the disappointments, and you even let me be petty without holding it against me. Thank you for always understanding the demands of the program and supporting me in meeting them

To close, I want to thank the specific educators and mentors who have helped and inspired me on my academic journey. Brett Hardin first inspired my interest in the social sciences as a teacher at Campbell High School, Danny Gulden facilitated my interest in social justice as a sponsor and youth minister at Sandy Springs Christian Church, and Eric Edmonds first taught me about empirical research as a professor and thesis advisor at Dartmouth College. Dana McCoy, Rekha Balu, and Robin Jacob have been role models since before I started grad school. They gave me opportunities to grow as a researcher, gave professional advice and support, and demonstrated what it meant to be brilliant and kind educational scholars. My committee members, Jimmy Kim, Luke Miratrix, and Marty West have been incredibly supportive and helpful throughout the PhD and in producing this dissertation. Marty, thank you for your pragmatism and helping me think strategically about how to accomplish my

professional goals. Luke, thank you for always being willing to jump into the mess of a methodological challenge with me, and for providing an example of following your passion. Jimmy, thank you for the thoughtfulness and care you have put into being my advisor and scaffolding my learning throughout this program. You have given me trust to carryout research on your behalf, opportunities to grow as a scholar, and funding to support my work. Most especially, you have extended me grace when I stumbled through, as you call them, temporary failures. To you all, thank you for pushing me intellectually and challenging me to improve the rigor and relevance of my research, while also giving me space to be vulnerable.

Dedication

This dissertation is dedicated to my late grandfather, James E. Armstrong Sr. (1921-2018). Though he never completed college, my grandfather was a life-long learner, passionate about education, and committed to supporting his children and grandchildren through higher education journeys.

Introduction

Intervention research in education seeks to understand whether, why, and how particular programs help students learn. Historically, this research has largely focused on estimating the overall average treatment effect or the average treatment effect for specific subpopulations of students, questions to which existing methods are well-suited. However, new evidence indicates that program effects can vary substantially (Weiss et al., 2017), and there is still much to be learned about which features, implementation contexts, or and student populations make complex programs effective. To shed light on this question, I investigate sources of treatment effect heterogeneity of early elementary literacy interventions both in home and school contexts. Leveraging creative experimental designs, each of these studies contributes to our broader understanding of treatment effect heterogeneity in causal education research.

Why heterogeneity matters

Treatment effect heterogeneity offers both challenges and opportunities for researchers and practitioners. This heterogeneity makes understanding program effects more difficult, because the average treatment effect may no longer be representative of the impact for the particular group of schools or students of interest (Imai et al., 2008). This is particularly important because we know that large-scale education research is often conducted in schools that are not representative of the nation as a whole (Fellers, 2017) and that the estimated effects of interventions in these samples are likely to differ meaningfully from those estimated in a random, nationally representative sample (Bell et al., 2016). Additionally, in a review of 16 large scale individual-randomized controlled trials in the social sciences, Weiss et al. (2017) systematically estimate the amount of cross-site variation in impacts and find that about a two thirds of studies demonstrate moderate-to-large amounts of heterogeneity in their effects.

Awareness of treatment effect heterogeneity is now widespread and a growing body of research leverages frameworks for studying this phenomenon (Duncan & Vandell, 2012; Schochet et al., 2014; Weiss et al., 2014).

Despite these challenges, understanding variation in impacts has the potential to improve education interventions. First, it can allow policymakers and practitioners to target interventions in contexts and populations where they are most effective and to avoid implementing them in situations where they are not. This targeted approach allows leaders to efficiently allocate scarce resources to maximally improve educational outcomes. Second, by understanding how components of interventions complement each other, interventions can be modified to increase their impact. Finally, a deep understanding of treatment effect heterogeneity will allow researchers to prepare for the challenges that often arise from scaling up interventions – maintaining fidelity when implementation is not supported by researchers and replicating the effects in new contexts. Altogether, these improvements would result in the amplification of estimated effects of existing interventions and a more efficient use of education resources.

The need to maximize the effectiveness and efficiency of educational interventions is particularly pressing in literacy. According to aggregate statistics, large swaths of students in the United States are failing to meet grade-level expectations of proficiency in reading (Nation's Report Card, 2015) despite the millions of dollars spent on interventions and research. Additionally, one meta-analysis finds that average treatment effects in literacy interventions are dwarfed in magnitude by average annual student learning gains (Hill et al., 2008; Scammacca et al., 2015). While more work will be needed to dramatically improve reading proficiency among American students, understanding treatment effect heterogeneity is a necessary first step toward helping the nation's struggling readers.

The included studies

In this dissertation, I present three studies, each of which explores a different facet through which impact variation might emerge. My approach is informed by the work of Weiss, Bloom, and Brock (2014), who identify three distinct sources of treatment-effect heterogeneity, which they label as the "3 C's – client characteristics, program context, and treatment-control contrast. Each of the three papers highlights how appropriate experimental design can result in deeper understandings of the effectiveness of educational programs and interventions. Highlighting the fundamental importance of literacy skills, the research questions in each paper are tightly connecting to the development and analysis of programs and policies to help students be successful readers.

The first two studies provide empirical demonstrations of how within-study variation in the treatment-control contrast yields variation in program impacts on reading. In the first study, which was jointly authored with Ethan Scherer, we explore variation in the effects of a parental text messaging intervention that is designed to increase student summer reading by comparing the effectiveness of different versions of the treatment. Drawing on behavioral science and home literacy theory, we use a factorial design to systematically compare three different characteristics of the text messaging campaign. The first component provides up-to-date personalized information in the text messages, compared with more generic messages. The second component provides parents with the opportunity to set a summer reading goal at the start of the messaging campaign, compared with parents who do not receive that option. Finally, we consider three different types of message framing. Some families receive messages that emphasize an instrumental, or skill-building, view of reading (Baker et al., 1997). Other families receive messages that emphasize an entertainment view of reading (Baker et al., 1997). Finally, some

families receive a combination of the two types of messages. We explore the differential effects of these components independently and in combination with one another on measures of family reading behaviors, student test scores, and parent and student experiences.

In the second study, I explore how the effects of a district-run, universal-access pre-K program vary based on the access of the control group to different counterfactual early childhood educational services. I leverage administrative records collected at kindergarten registration to describe the distribution of early childhood education across the district and in specific administrative regions. Then, using enrollment records and pre-K program waitlists, I construct two analytic samples of families who expressed interest in attending the district pre-K program. By limiting the sample to oversubscribed programs that held lotteries to determine enrollment, I estimate plausible quasi-experimental estimates of the effect of attending the district pre-K program on measures of kindergarten readiness. I then compare these estimates across administrative regions where the control group has different experiences and kindergarten readiness.

The final study is a methodological investigation of an emerging experimental design in education research: Sequential Multiple Assignment Randomized Trials (SMARTs, Almirall et al., 2014). SMARTs are used to develop, refine, and test adaptive interventions, where the nature of the treatment can be changed for the second phase of the intervention to better address the needs of the sample. Using the potential outcomes framework and the concept of principal stratification, I define several endogenous subgroups determined by how individuals will respond to the different treatments available in the first phase of the adaptive intervention. I also define the corresponding estimands related to average treatment effects. Then, using simulations, I

show under what conditions common analytic techniques allow researchers to recover these estimands and to estimate the magnitude of bias when they fail to do so.

Learning what works for whom and in what contexts is a central question confronting researchers and practitioners. Through this dissertation, I hope to advance our understanding of how research design can reveal treatment effect heterogeneity in education.

Paper 1: Using a Factorial Design to Maximize the Effectiveness of a Parental Text Messaging Intervention[1]

Background

Both policymakers and researchers are increasingly using insights from behavioral sciences to enhance parental engagement and improve academic behaviors such as attendance, as well as academic skills like reading and math. One subset of these interventions provides parents with younger children timely and actionable text messages, many of which contain reminders of home literacy activities. These interventions target several potential behavioral levers to increase parents' engagement. First, receiving a text message could refocus parents' attention to educational actions (Cortes et al., 2019; Cortes et al., 2021; Doss et al., 2019; Hurwitz et al., 2015; Kim et al., 2019; Kraft and Monti-Nussbaum, 2017; Mayer et al. 2018; Smythe-Leistico & Page, 2018; York et al., 2019). Second, text messages can help parents navigate the challenge of short-term, immediate-cost behaviors (like reading to your child every day) whose benefits do not manifest until the future. Mayer et al. (2018) find setting goals and monitoring parents progress towards these goals mitigated these self-control concerns. Third, a subset of the studies attempts to reduce the complexity, or cognitive load, of parenting by breaking down parental education activities into several discrete components (Hurwitz et al., 2015; Kraft and Monti-Nussbaum, 2017; Cabell et al., 2019; Doss et al., 2019; Kim et al., 2019; York et al., 2019). They provide a combination of literacy facts or resources, a specific activity or practice, and extension activities that parents can implement without additional preparation. Finally, informational messages attempt to facilitate parent monitoring of their children's academic performance. For

[1] Co-authored with Ethan Scherer

example, it is difficult to track and monitor a child's cumulative absences, and parents often underestimate how much school their child has missed (Smythe-Leistico & Page, 2018). One method to address these concerns is providing parents with up-to-date information on their child's performance to correct these biased beliefs.[2]

Across these studies, most find positive effects on behavioral measures of parent or child reading, such as the amount of reading, or parental involvement, often in the magnitude of 0.15-0.3 standard deviations (Doss et al., 2019; Hurwitz et al., 2015; Kraft & Monti-Nussbaum, 2017; York et al., 2019) and Smythe-Leistico and Page (2018) find an 11-percentage point decrease in absenteeism. While it is clear from these studies that messaging interventions can change parental behavior, there is less clear evidence that they can influence more distal outcomes like student reading skills, as measured by test scores. While a few studies find positive effects of around 0.10-0.18 standard deviations (Doss et al., 2019; Kim et al., 2019; York et al., 2019) several find no significant differences or even mixed results for all children (Cortes et al., 2019; Cortes et al., 2021; Kraft & Monti-Nussbaum, 2017) and some do not examine test scores at all (Hurwitz et al., 2015; Mayer et al., 2018; Smythe-Leistico & Page, 2018). This variation could be related to different samples and context, but also suggests the importance of the specific behavioral and content features of the message.

Furthermore, while the existing research increasingly indicates the focus of the message content is important, interventions often combine message types or other components to maximize the intervention's effectiveness. In Doss et al. (2019), for example, the messages combine facts and information about reading with specific activities for parents to complete with

[2] Other studies have used informational messaging campaigns administered through a non-texting medium to reduce absenteeism among students (Rogers & Feller, 2018; Rogers et al., 2017; Robinson et al., 2018).

their children in both arms of the intervention. Similarly, Mayer et al. (2018) combine their text messaging with regular check-in meetings with the program provider. Thus, while it is clear that the characteristics of the intervention matter, it is hard to assess whether individual components could work as well as a cocktail of messages or if the synergies between the components drive the effects. Unfortunately, typical randomized controlled trials, by design, are ill-suited to efficiently unpack the effectiveness of individual intervention components and their combination to answer the question of what works, for whom, and why.

While texting interventions demonstrate reduction in specific behavioral barriers, less is known about how messaging interventions inform parent's domain-specific beliefs. For example, all the previously mentioned literacy texting interventions focus the content of their messages on building literacy skills, a framework aligned with the instrumental view of reading that emphasizes the value of skill-building for future success (Baker et al., 1997). However, literacy scholars document that while some parents articulate this instrumental view, others emphasize an entertainment view, emphasizing reading for pleasure and enjoyment (Baker et al., 1997). A series of small-sample studies finds that these parental beliefs differentially predict features of their children's reading experiences. Parents tend to promote types of reading activities for their children that align with their beliefs (Lynch et al., 2006; Sonnenschein et. al, 1996; Sonnenschein et. al, 1997). Additionally, parents who endorse the entertainment view have children who score higher on measures of reading enjoyment and motivation (Baker et al., 1997; Baker & Scher, 2002). Correlational research indicates that students with exposure to the entertainment view tend to be better readers than similar students from homes who emphasize the instrumental value (Baker et al., 2001; Sonnenschein et al., 1997; Sonnenschein et al., 2000). Less is known, however, about whether these differences in views and activities are causally related to student

outcomes, and how underlying student motivation may be influenced by the introduction of different views. Causally investigating how literacy parental texting intervention effect could vary based upon promoting particular beliefs remains a critically unexplored area of literacy parental texting research.

The prior literature raises important unanswered questions for parental text messaging interventions about the specific behavioral and content features that make them most effective. In particular, past work has suggested the importance of both data-driven personalized message content and goal setting, but no work to our knowledge, has assessed these levers at scale individually or in combination. Furthermore, while the messaging to parents about reading beliefs has been pushed as a promising area of intervention, it has never been explored causally and never been combined with other behavioral interventions. Our research seeks to understand what types of parental text messages are most effective, by investigating two primary research questions:

1. What components and combination of components of parental text messages affect parent behavior?
2. Under what conditions do the effects on parental behavior transfer to student reading outcomes?

We address these questions by implementing a parental text message intervention to understand the individual and combined effects of individual message components on family reading behaviors and student reading comprehension. Specifically, we examine the effectiveness of three distinct components: personalizing messages with up-to-date information; the message's framing of reading as building skills, for entertainment or both; and setting reading goals at the beginning of the summer. We test these components of the text messaging

campaign using a 2x2x3 factorial experiment with approximately 5,000 elementary school students in a large school district in the southeast United States. A factorial design is particularly well-suited to study an intervention with multiple distinct components that could be individually included or excluded, because it allows for the estimation of each component's individual effects as well as interaction effects of multiple components (Collins et al., 2014; Somers et al., 2014).

A unique feature of our intervention is our ability to both directly and indirectly measure parental reading behaviors in response to our text messages. All participants in our study had access to a summer reading program and accompanying educational reading app. Over sixty percent of our text messages encourage parents to have their child login and use the reading app. Given that we provide login passcodes only to the parents, we can observe parental engagement with summer reading based on app usage data. Similar to prior studies, we also directly survey a subsample of parents.

We find that in comparison to generic messages, messages containing personalized information increase parental reading supports of reading and student reading performance in the fall. Students of parents who received text messages were three percentage points more likely to use the app. These students also performed 0.03 standard deviations better on a standardized reading assessment used formatively as well as .02 standard deviations better on high stakes beginning of grade exams, though the latter was not statistically significant. Levering the interaction effects within our factorial design, we also find that the view of reading emphasized in the messages can enhance or detract from these information effects, with a combination of reading values magnifying the positive effects of personalized information when compared to personalized messages that focused on instrumental values alone. We also explore potential mechanisms for the increase in reading comprehension test scores using our rich set of student-

level outcomes. An important hypothesis from the view-of-reading literature posits that emphasizing entertainment values will increase a child's motivation, and ultimately improve reading comprehension. We directly test changes in reading motivation using student self-reports from both a paper survey, and among the students who logged into the app, responses to reading motivation questions directly embedded into the app but find no evidence to support this claim.

We structure the remainder of the paper in the following way: first, we provide a brief description of our intervention, data, measures, and methods used in our analysis; next, we present our main results and an investigation of potential mechanisms; we conclude with a discussion of our findings and their implications for texting interventions for parents.

The Intervention

Context and Study Eligibility

Students were recruited to participate in this study through their participation in a large, multi-school randomized control trial (RCT) of a curricular intervention. All first and second grade students in the participating schools were invited to participate in both the curriculum and texting studies through an active consent process. For the curriculum study, either first or second grade students at each school were assigned to receive the MORE (Model of Reading Engagement) intervention curriculum, with the other grade serving as a control group, who received the business-as-usual curriculum. This intervention consisted of a series of science- and social studies-themed lessons in the spring semester and the students' choices of 10 hardcopy books related to the same science and social studies topics presented in the classroom lessons.

To be eligible for our text messaging study, the research team had to identify a valid cell phone number for a parent or guardian of each student. Phone numbers were either provided directly on the consent form or by the school district administrative records. Once a cell phone

number was validated, the families were enrolled in the text messaging study. Participating families were then randomized into conditions that received different versions of the texting intervention, including a small group that received no text messages.

At the end of the school year, all students in both the curriculum study and texting study received access to an educational reading app called MORE@Home, which contained six digital books as well as a series of reading-related activities matched to each book and broadly leveled according to the child's end-of-year reading ability. For those students in the MORE group of the curriculum study, their MORE@Home accounts also provided access to leveled reading activities for each of the 10 books they had selected. Overall, use of the app was relatively low across the sample. Among our pure control group of families, only 16% of student accounts were ever activated by their parents.

Text Messaging Intervention

The purpose of the text messaging intervention was to increase parental engagement with their children in summer reading activities and improve student learning. In line with recent research on the effect of frequency and timing of text message interventions (Cortes et al., 2019; Cortes et al., 2021), families received text messages twice weekly over 9 weeks of summer vacation, with one message occurring earlier in the week and one message closer to the weekend. Messages were sent in either English or Spanish, based on the student's home language in district administrative records.[3] Messages were all sent through the Twilio messaging platform which was accessed through our sample database. This approach had several advantages: 1) message times were scheduled in advance; 2) messages were linked to user profiles containing

[3] Our students had several other home languages designated, in addition to English and Spanish, though none represented more than (1%) of students. Families whose home language was neither English nor Spanish received English messages due to resource constraints.

other research information from the curriculum study and educational app database; 3) families could easily opt-out of messages, preserving consent; and 4) parental responses to text messages were logged on the user profiles. Some text messages were designed specifically to promote usage of the educational reading app, while others encouraged a wider variety of reading activities. Each message contained a single topic, which generally covered one of the three larger themes: 1) reminders to engage in summer reading activities; 2) providing information about summer reading resources (including the educational app); and 3) monitoring progress throughout the summer.

Differentiating Messages

To explore how text messaging features differentially influence parental and student engagement with a summer reading intervention, we differentiated the specific wording of each text message topic according to three separate factors that have shown promise in prior research: updated personalized information to correct parent informational misbeliefs (Smythe-Leistico & Page, 2018), goal-setting to reinforce immediate action (Mayer et al., 2018), and framing different views of reading to promote specific parental beliefs (Baker et al., 2001; Sonnenschein et. al, 1997; Sonnenschein et al., 2000).

In the personalized information factor, some families received text messages that include student-specific information within the message. Some examples of the information that could be included were the specific books the student had access to in the app, whether or not a student had logged into the educational app yet, or which books' activities the students had accessed on the app. Because the text messaging was integrated with the app's backend database, each student's information was updated continuously, ensuring that the messages reflected the students' most recent status. Families not in the personalization condition received more generic

messages, but they still referred to individual students by name.

For the goal-setting factor, some families were invited to set a summer reading goal at the beginning of the intervention, with later messages periodically checking-in on their progress towards that goal. We designed the goal setting to be a light-touch, low-cost, scalable version of other effective goal-setting studies (e.g., Mayer et al., 2018; Oreopoulos et al., 2020), with parents being asked not only to identify a goal but to make a plan for reaching it in the face of obstacles (Oettingen & Reininger, 2016). However, without an in-person goal setting session or subsequent individualized follow-up, most families (more than 95% of the goal-setting condition) failed to complete the goal-setting exercise. For the families who did set a reading goal, check-in messages would explicitly refer to their individual goal. Other messages referred to "your summer reading goal."

Finally, the view-of-reading value's condition created three distinct groups who received differently framed messages over the course of the campaign. The instrumental-only-view condition emphasized reading as a process by which students develop specific skills important for future success. The entertainment-only-view condition emphasized reading as an enjoyable and fun activity. The combination-view received a balanced combination of entertainment- and instrumental-framed messages over the course of the summer. Importantly, however, this combination-view received the same total number of messages as the entertainment-only and instrumental-only conditions.

Not all messages included components based on the levels of all 3 factors. A single message could meaningfully differentiate between conditions for just goal-setting factor, just the personalization factor, just the view-of-reading factor, or across any combination of those factors; in our text messages, 40% were relevant with regards to families' goal-setting condition,

47% differed with regards to personalization, and 85% were framed in reference to a specific view of reading. To ensure that the intended conditions were salient to families, we recruited colleagues to review example text message versions and provide feedback on whether the messages clearly contained information aligned to specific conditions. In cases where colleagues' perspectives differed from our intention, we revised the message versions to increase or decrease the salience of a specific condition. A comprehensive list of message themes as well as example variations of a given message can be found in Appendix A.

Methods

Sample

This study includes 5,172 rising second and third grade students, from 4,993 families, attending thirty elementary schools in a large school district in the southeast. Demographic characteristics are presented in the first column of Table 1.1. Close to forty percent of the students are African American and an additional thirty percent are Hispanic. Approximately twenty percent are white, and ten percent are Asian. Almost a quarter of students were receiving English-learner services. Our sample contains socioeconomic diversity but contains a larger proportion of students in low-SES neighborhoods relative to the district as a whole.

Research Design

We use a factorial experiment to compare the differential effectiveness of text messaging components. Traditional randomized controlled trials are only able to compare two treatment arms at a time, so investigating three potential mechanisms would require multiple experiments, or a multi-arm RCT with a prohibitively large sample size. A factorial design, however, is

Table 1.1.

Baseline characteristics and balance checks

	Full Sample Mean	Personalization vs. Not Difference	Entertainment vs. Instrumental Difference	Combination vs. Instrumental Difference	Goal Setting vs. Not Difference
White (%)	0.184	0.004	-0.000	-0.008	-0.001
	(0.388)	(0.009)	(0.012)	(0.010)	(0.009)
Black (%)	0.386	0.008	-0.023	-0.011	0.008
	(0.487)	(0.012)	(0.018)	(0.017)	(0.012)
Hispanic (%)	0.317	-0.008	0.022	0.001	-0.004
	(0.465)	(0.011)	(0.017)	(0.017)	(0.013)
Asian (%)	0.079	-0.001	0.008	0.023*	0.003
	(0.270)	(0.008)	(0.012)	(0.011)	(0.008)
Limted English Proficiency	0.226	-0.016	0.014	-0.000	0.013
	(0.418)	(0.010)	(0.012)	(0.015)	(0.012)
Low SES	0.405	0.017+	0.004	-0.004	-0.004
	(0.491)	(0.010)	(0.012)	(0.012)	(0.009)
Med SES	0.385	-0.000	0.002	0.011	0.004
	(0.487)	(0.010)	(0.015)	(0.014)	(0.011)
High SES	0.205	-0.018*	-0.009	-0.010+	-0.001
	(0.404)	(0.009)	(0.009)	(0.006)	(0.010)
Baseline ELA Scores	-0.000	0.040+	0.001	-0.014	-0.022
	(0.988)	(0.024)	(0.035)	(0.035)	(0.025)
Baseline Math Scores	-0.000	0.052+	0.001	0.005	-0.042
	(0.988)	(0.027)	(0.033)	(0.038)	(0.028)
N	5175	4678	3119	3128	4678

Source: district administrative records

Notes: Each row represents a separate regression. Point estimates reflect the condition-reference group differences derived from a model that includes indicators for the randomization block. Robust standard errors clustered at the school-grade level (in parentheses).
+p<0.10, *p<0.05, **p<0.01, ***p<0.001

particularly well-suited to study an intervention with multiple distinct components. In a factorial design, each intervention component is treated as its own factor, with different levels representing the treatment assignment. Each unit is randomized to a level for each factor independently. This design has the benefit of allowing the researcher to test the average main effect of each intervention component, as well as interaction effects of intervention components (Collins et al., 2014; Somers et al., 2014). It is thus an appropriate design to address how the multiple levers targeted in texting messaging interventions contribute to an intervention. We use a full-factorial design, in which every factor is fully interacted with the other factors. The goal-setting and personalized information factors each have two levels (on, off), and the view-of-reading factor has three levels (instrumental view only, entertainment view only, both views presented), resulting in 12 different treatment combinations. Additionally, we assigned a small portion of the sample to a pure control condition, not receiving any text messages, as shown in Table 1.2. Separating out this pure control provides a business-as-usual condition to use as a benchmark for the magnitude of the factorial differences, but our research questions focus exclusively on the relative effectiveness of the different text message components, not the effects of text messaging compared to no messaging.

To account for the presence of siblings in the sample, which could result in spillover and confusion for parents receiving two types of messages, random assignment to conditions occurred at the family-level. To improve precision and reduce the minimum detectable effect size, the sample was blocked at the school-by-grade level, the unit of treatment from the larger RCT, because average student reading levels and implementation fidelity of the larger intervention vary across schools. Within these blocks, the 4,993 families (5,172 children) with valid cell phone numbers were assigned to one of the 13 conditions.

Table 1.2

Treatment Conditions in Factorial Design with Effect Coding

Group	Received Text Messages	Personalization Condition	Value	Goals Condition	Value	Reading Value Condition	Value (Entertain)	Value (Both)
1	No		.		.		.	.
2	Yes	No Personalization	-1	No Goals	-1	Instrumental Only	-1	-1
3	Yes	No Personalization	-1	No Goals	-1	Entertainment Only	1	0
4	Yes	No Personalization	-1	No Goals	-1	Both	0	1
5	Yes	Personalization	1	No Goals	-1	Instrumental Only	-1	-1
6	Yes	Personalization	1	No Goals	-1	Entertainment Only	1	0
7	No	Personalization	1	No Goals	-1	Both	0	1
8	Yes	No Personalization	-1	Goals	1	Instrumental Only	-1	-1
9	Yes	No Personalization	-1	Goals	1	Entertainment Only	1	0
10	Yes	No Personalization	-1	Goals	1	Both	0	1
11	Yes	Personalization	1	Goals	1	Instrumental Only	-1	-1
12	Yes	Personalization	1	Goals	1	Entertainment Only	1	0
13	Yes	Personalization	1	Goals	1	Both	0	1

A series of balance checks are presented in the remaining columns of Table 1.1, comparing our sample across each factor of the intervention. Overall, we find few differences between experimental groups on baseline demographic characteristics or academic performance, and none that are statistically significant after applying a multiple hypothesis correction to account for the number of student characteristics compared (Benjamini & Hochberg, 1995). The similarities between groups reflects a successful randomization process.

Data Sources

Baseline data from district administrative records of the pre-intervention year (2018-2019) is available for our entire study sample. These measures include enrollment information, student demographics, and reading and math test scores. The school district where this intervention took place does not collect student-level measures of socio-economic status; however, student neighborhoods, as determined by their census block-group, are categorized as

being low-, middle-, or high-SES communities. Student-level outcome data from Fall 2019 is also provided by the district. We use the Measure of Academic Progress (MAP) RIT score in literacy (Northwest Evaluation Association, 2011) as a primary outcome, which measures foundational literacy skills. A second academic outcome, available for the rising third-grade cohort, is the Beginning of Grade (BOG) assessment, a statewide measure designed to measure student's progress towards proficiency on the End of Grade (EOG) accountability assessment at the end of third grade. The BOG is also used for school accountability growth and for the identification of exceptional teachers. Both assessment scores are standardized within grade to provide outcomes in standard-deviation units.

We assess the effect of the intervention on parents' behaviors and beliefs with two sources. First, we track family use of the MORE@Home app. Because 60% of the text messages parents received were related to the educational app, usage statistics reflect whether parents changed their own and their children's behavior in response to messages. Specifically, we are interested in whether parents logged their students into the app, the total number of books they accessed on the app, and the total number of minutes they spent engaged with the app. For families who never logged into the app, total books and total minutes were recorded as zeros. We also collected self-reported outcomes from a subsample of parents who were randomly invited to complete a parent survey. The survey included questions about their summer reading activities outside of the educational app and their perceptions of the text messages they received.

To explore student-level mechanisms and provide more information about students' summer reading experiences, we collected two waves of student surveys as well as qualitative measures of the students' use of the educational app. In the spring prior to the summer texting intervention, we measured student motivation using the Me and My Reading Profile (MMRP,

Marinak et al., 2015). In the fall, we administered a student survey that included the MMRP and additional questions about students reading behaviors over the summer. The fall survey was administered in the thirty study schools and thus does not cover either the students who moved out of the district or those who moved to a non-study school within the district. For students whose parents logged them into the app, we measured the percentage of activities that they completed correctly. We also asked periodically whether they enjoyed the app, felt like a good reader, and found the app activities challenging. Because our sample is slightly different for each set of outcomes, we test for differential attrition rates based on the different factors of our treatment. We find no significant differences in retention rates by condition for test scores, parent and student surveys, and objective app metrics. The sample for the qualitative app experience was over-represented by students in the personalization condition.

Empirical Strategy

We analyze the twelve conditions in our factorial experiment using standard regression techniques. Recent work by Muralidharan et al. (2020) highlights the importance of using a model that includes all treatments from the factorial experiment and their interactions. Reviewing recent factorial studies in economics, the authors find that leaving out treatment interaction terms yield incorrect inferences of main effects if the interaction effects are non-zero or if model-selection is determined after looking at the magnitude and significance of these interaction effects. To avoid both of these concerns, and to allow for concurrent interpretation of main effects and interaction effects of our different factors, we use the following model, where Y_{ij} represents the outcome for individual i in randomization block j:

$$Y_{ij} = \beta_1 Pers_{ij} + \beta_2 EntView_{ij} + \beta_3 BothView_{ij} + \beta_4 Goals_{ij} + \beta_5 PersxEntView_{ij}$$
$$+ \beta_6 PersxBothView_{ij} + \beta_7 PersxGoals_{ij} + \beta_8 GoalsxEntView_{ij}$$
$$+ \beta_9 GoalsxBothView_{ij} + \beta_{10} PersxEntViewxGoals_{ij}$$
$$+ \beta_{11} PersxBothViewxGoals_{ij} + \Gamma X_{ij} + \phi + \varepsilon_{ij}$$

The model also includes a vector of covariates X_{ij}, including student demographics and pre-test math and reading scores, as well as a set of fixed effects, ϕ, representing the randomization blocks. Standard errors are clustered at the school-grade level as recommended by Athey and Imbens (2017), to account for the unit of blocking and the correlation of residuals within those blocks.

The main effects (captured in β_1 through β_4) provide information about the average effect of each text messaging factor, and the two-way interactions (captured in β_5 through β_9) tell us whether the effects of one factor depend on the levels of the other factors. The three-level view-of-reading factor has been separated into two variables, EntView (for entertainment only) and BothView (for a combination of values), with the instrumental view-of-reading condition serving as the reference category for both of those variables. The three-way interaction terms (β_{10} and β_{11}) are included to allow us to estimate the main effects and two-way interactions concurrently; given their difficulty of interpretation, they are not parameters of interest for this paper.

To facilitate interpretation of effects, treatment variables are coded using effect coding with variables taking on values of -1 or 1 (Kugler et al., 2012). This parameterization of the treatments highlights the benefits from using effect coding. Because our sample is evenly divided across conditions, this parameterization allows both the main effects (the difference between levels of each factor or marginal effects) and the interaction effects (the additional effect of receiving a particular combination of factors) to be estimated concurrently (Hardy, 1993).

However, because level differences require the treatment indicators moving from -1 to 1 instead of from 0 to 1 as in the standard dummy coding, we must multiply all coefficients (and their standard errors) by 2. Thus, we can interpret the main effect β parameters as follows:

- The average effect of receiving personalized information (vs. not) is $2\beta_1$.
- The average effect of framing messages with the entertainment views of reading (compared to only instrumental-only view) is $2\beta_2$.
- The average effect of framing both views of reading (compared to only instrumental-only view) is $2\beta_3$.
- The average effect of setting goals (vs. not, as if in a traditional RCT) is $2\beta_4$.

Interaction terms are interpreted as normal once they have been scaled up: the additional effect of one factor in the presence of another factor. In all tables in this paper, we have already adjusted the point estimates and standard errors for ease of interpretation. Because we present multiple outcomes in each domain, we also test our confirmatory results to the sensitivity of false discoveries, using the Benjamini-Hochberg procedure with a false discovery rate (FDR) set to 0.05 (Benjamini & Hochberg, 1995) by outcome domain.

Results

Effects on Parental Behaviors and Beliefs

We first consider how the different types of text messages affect parental behaviors and beliefs in Table 1.3. Panel A presents the main effects of each component, and Panel B presents the interaction effects, but for each column, both panels come from the same fitted model. Our primary measures of behavioral change in parents are captured in their use of the educational reading app with their children. For all outcomes, the probability of ever logging into the app, the total minutes spent on the app and the number of books completed, the personalized messages

were significantly more effective than non-personalized messages. Families receiving personalized messages were three percentage points more likely to use the app (ES=0.08), spent about an extra 1.6 minutes using the app (ES = 0.11) and completed an additional 0.7 books worth of activities (ES=0.12). These effects remain statistically significant after multiple-hypothesis corrections (Benjamini-Hochberg, 1995).

Table 1.3

Differential Effects of Text Messaging Components on Parent Outcomes

	Observed reading behaviors in app			Self-reported behaviors & beliefs		
	Ever logged in (%)	Minutes on app	Books completed	Frequency of reading activities	Found texts helpful (%)	Would recommend texts to others (%)
Panel A - Main Effects						
Personalization vs. Not	0.028*	1.626**	0.681**	0.046	0.023	0.013
	(0.012)	(0.560)	(0.254)	(0.144)	(0.163)	(0.074)
Entertainment vs. Instrumental	0.004	-0.386	0.033	-0.273+	-0.090	-0.112
	(0.020)	(1.016)	(0.434)	(0.160)	(0.272)	(0.072)
Combination vs. Instrumental	-0.020	-0.179	-0.345	0.249	0.360	0.068
	(0.020)	(0.891)	(0.397)	(0.191)	(0.243)	(0.093)
Goals vs. Not	-0.013	-0.245	-0.187	0.131	0.318+	0.027
	(0.013)	(0.675)	(0.287)	(0.156)	(0.169)	(0.072)
Panel B - Two-Way Interaction Effects						
Personalization x Entertainment	-0.013	-0.282	-0.149	0.191	0.211	0.012
	(0.019)	(1.010)	(0.435)	(0.150)	(0.234)	(0.081)
Personalization x Combination	0.006	-0.116	-0.187	-0.255	-0.097	-0.111
	(0.016)	(0.948)	(0.391)	(0.183)	(0.225)	(0.088)
Personalization x Goals	0.014	0.365	0.138	-0.044	-0.243	-0.051
	(0.012)	(0.587)	(0.234)	(0.128)	(0.176)	(0.057)
Entertainment x Goals	-0.021	-0.947	-0.464	-0.089	0.058	0.021
	(0.017)	(0.855)	(0.337)	(0.250)	(0.178)	(0.066)
Combination x Goals	0.041*	1.858*	0.686+	0.129	0.280	-0.003
	(0.020)	(0.945)	(0.395)	(0.176)	(0.196)	(0.088)
N	4678	4678	4678	319	266	263

Source: app-use records, survey of parent subsample, district administrative records

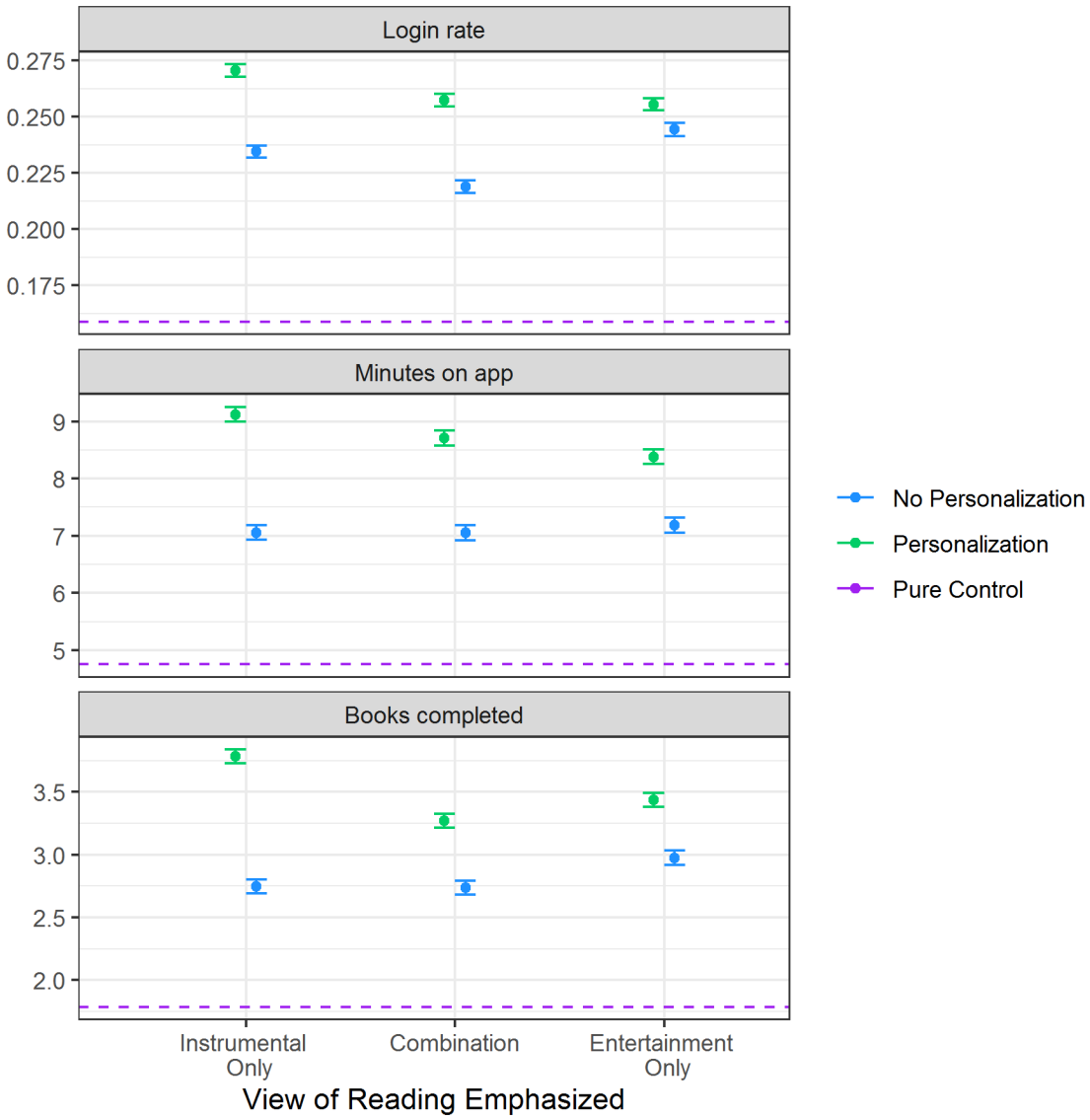
Notes: Point estimates derived from effect-coded regressions that include all treatment factors and their interactions, as well as the following covariates: gender, race/ethnicity, participation in gifted program, participation in Special Education, English learner status, neighborhood SES, language of text messages, baseline reading and math scores, and indicators of the randomization block. Robust standard errors clustered at the school-grade level (in parentheses).

+p<0.10, *p<0.05, **p<0.01, ***p<0.001

To facilitate this interpretation, Figure 1.1 presents model-based predicted outcomes for our sample and their standard errors, grouped by personalization and the view of reading,

Figure 1.1

Predicted Reading Behaviors, by Personalization and View-of-Reading Conditions



Source: app use data, district administrative records

Notes: Predicted values and standard errors for the six groups are calculated from model parameters and aggregated by the personalization and view-of-reading conditions, collapsing across the values condition. The pure control value is observed directly from the data.

and pooled across the goal-setting condition. The personalized text results are in green and the generic texts are in blue, and the horizontal groupings represent the view-of-reading conditions. From these, we can more easily identify common trends that are not immediately apparent from the model output. This figure highlights that personalization improved app use for all three view-of-reading conditions, but that compared to our control condition, both generic and personalized messages were effective at increasing parental engagement with reading activities. These results are consistent with prior work on absenteeism using postcards that providing personalized information to parents debiases their belief when information is hard to obtain (Robinson et al., 2018; Rogers & Feller, 2018).

We also consider self-reported behaviors and beliefs from the subsample of parents who responded to our parent survey. Because the survey sample is relatively small, we only examine the main effects of each factor. Among this group, parents receiving personalized activities also report engaging in more frequent reading activities with their students (ES = 0.05), but this difference is not significant. These results demonstrate that personalized text messages caused parents to login and use the reading application more.

For the other two factors, while the point estimates are generally not significant, the reading behaviors observed in the app and the self-reported behaviors trend in the opposite directions. For example, there are only small, insignificant main effects on directly observed app usage from changing the view-of-reading and providing parents with the opportunity to set a summer reading goal. However, we do see larger differences on self-reported reading activities, some of which are significant at the 0.10 level. For example, relative to those receiving instrumentally framed messages, receiving only entertainment-framed messages decreased the frequency of reading by 0.27 standard deviations ($p<0.10$), whereas receiving a combination of

framing increased the frequency of reading by 0.25 standard deviations ($p>0.10$).

The view-of-reading components and goal setting also affect parental beliefs about the intervention. Providing a goal-setting opportunity increased the likelihood that parents found the text messages useful by 32 percentage points ($p<0.10$), as did receiving messages framed around both views of reading compared to only the instrumental view (by 36 percentage points, though not statistically significant). These features also made parents slightly more likely to recommend text messages to other parents, but these differences are not significant. We speculate that the divergence of app and self-reported results could be because the content of these messages change parental activities beyond the limited behaviors that we can observe in the app.

In Panel B, we see that goal setting and the view of reading have significant interaction effects. The effectiveness of a combination of values increases app usage when combined with goal setting, even though the individual components were not significant on their own. The combined effect of goal setting and a combination of view-of-reading was significantly more than either of the individual effects of app usage: 4 percentage points higher login rate, 1.8 additional minutes on the app, and an addition 0.7 books completed. These effects are similar in magnitude to the main effects of personalization described above but need to be considered in the context of the non-significant main effects of the goal setting and view-of-reading factors. Thus, goal setting alone does not seem to have an effect on directly observed app usage, but the combination of goal setting and changing the view of reading does affect these outcomes.

Effects on Student Reading Performance

To understand whether these changes in parental behaviors and beliefs translate into effects for their students, we present effects on reading scores in Table 1.4, following the same panel structure to consider both main effects and interaction effects simultaneously.

Table 1.4

Differential Effects of Text Messaging Components on Student Test Scores

	MAP	Beginning of Grade
Panel A - Main Effects		
Personalization vs. Not	0.034*	0.019
	(0.016)	(0.021)
Entertainment vs. Instrumental	0.028	0.021
	(0.025)	(0.033)
Combination vs. Instrumental	0.026	-0.008
	(0.022)	(0.032)
Goals vs. Not	0.007	0.007
	(0.016)	(0.027)
Panel B - Two-Way Interaction Effects		
Personalization x Entertainment	-0.011	-0.078*
	(0.028)	(0.037)
Personalization x Combination	0.025	0.106***
	(0.027)	(0.032)
Personalization x Goals	-0.005	-0.010
	(0.016)	(0.020)
Entertainment x Goals	-0.009	0.016
	(0.024)	(0.040)
Combination x Goals	-0.010	-0.029
	(0.021)	(0.042)
N	3961	2039

Source: district administrative records

Notes: Point estimates derived from effect-coded regressions that include all treatment factors and their interactions, as well as the following covariates: gender, race/ethnicity, participation in gifted program, participation in Special Education, English learner status, neighborhood SES, language of text messages, baseline reading and math scores, and indicators of the randomization block. Robust standard errors clustered at the school-grade level (in parentheses).

+p<0.10, *p<0.05, **p<0.01, ***p<0.001

We see evidence of transfer from the behavioral effects of personalized information to student test score outcomes. Personalized text messages significantly improve Fall MAP scores by 0.03 standard deviations. While not significant, the point estimate for the effect of

personalization on Beginning of Grade (BOG) scores is also positive, though slightly smaller than the effect on MAP (ES = 0.02). The MAP effect remains marginally significant after correcting for multiple hypotheses. Despite evidence that goal setting and view-of-reading differences affected parental experiences with the intervention, we find no significant main effects for changing either of these components on student test scores. We interpret the positive effects of personalization as transfer of the increased parental reading behaviors with the app because many of the app activities were designed to support the specific foundational literacy skills measured by the MAP assessment. The BOG effects represent farther transfer, which may be why they are smaller and less precise. The differences in effects between the two outcomes are not due to differences in sample; in sensitivity analyses we find similar effects on the MAP when we limit the sample to rising third graders who also took the BOG.

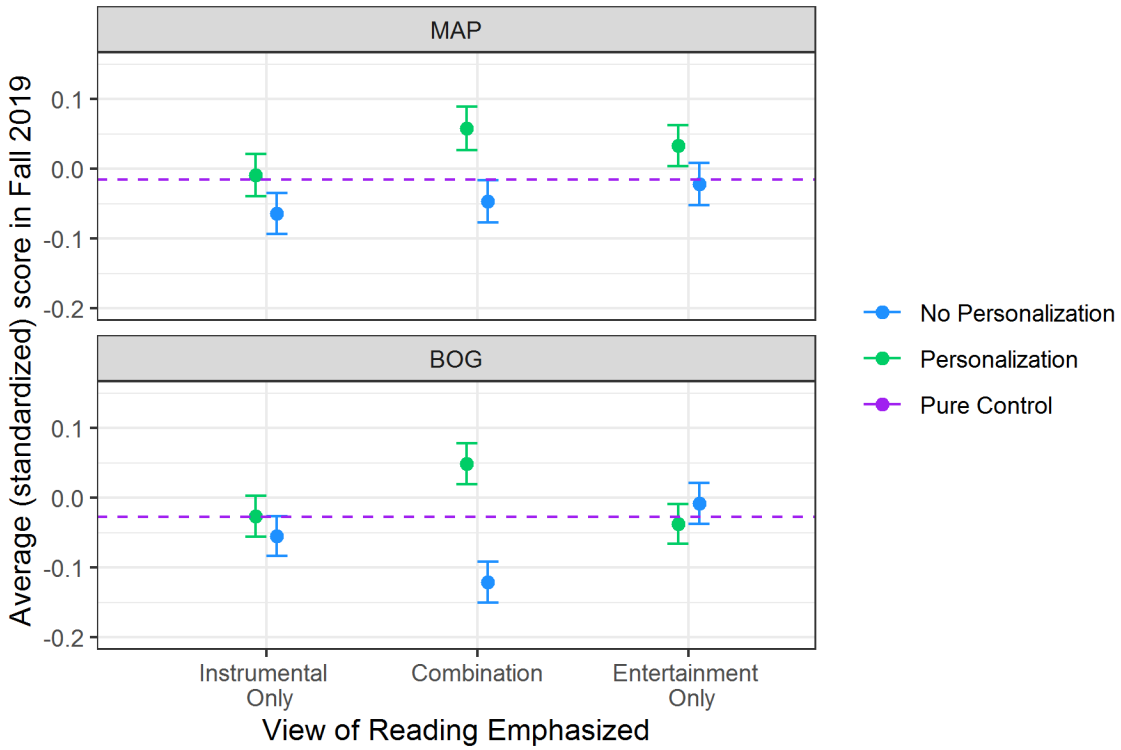
Panel B shows the interaction effects when we consider whether the observed main effects of each factor depend on the levels of the other components. Though the point estimates are only significant for Fall BOG scores, we see a similar pattern across both outcomes: personalization with the entertainment-only view is slightly less effective than personalization with instrumental only view (for the BOG, ES = -0.077, $p<0.05$), but personalization with both reading views presented is more effective than personalization with instrumental only values emphasized (for the BOG, ES = 0.104, $p<0.001$). Taken together with the positive interaction effects of the combined view and goal setting on parental reading behaviors, these effects on student test scores are further evidence that the view-of-reading framing of messages has the potential to enhance or detract from the effects of other factors.

Figure 1.2 shows these results visually and replicates the layout of Figure 1.1, where we plot the average predicted test score outcomes and their standard errors for each of the six groups

determined by levels of personalization and the view-of-reading emphasized (collapsing across levels of goal setting). From the figure it is easy to see that personalization outperforms no personalization because the green markers are almost always above the blue markers across MAP and BOG. In addition, focusing on the blue markers, absent personalization of text messages, students perform about 0.05 standard deviations better on fall reading assessments when they receive text messages emphasizing the entertainment view as opposed the instrumental view of reading. And while we saw from the output that personalization improves

Figure 1.2

Predicted Test Scores, by Personalization and View-of-Reading Conditions



Source: app use data, district administrative records
Notes: Predicted values and standard errors for the six groups are calculated from model parameters and aggregated by the personalization and view-of-reading conditions, collapsing across the values condition. The pure control value is observed directly from the data.

test scores on average, Figure 1.2 highlights that this effect is two to three times larger for students receiving a combination of reading values than for those receiving a single type of reading value, reflecting the significant interaction effects. Figure 1.2 also includes the standardized outcomes for our pure control group as a reference. The observed average control outcome highlights that our effects are not simply due to the mode of our intervention – the specific content and framing of the message matters. While the most effective types of text messages are helpful for student reading, less-effective messages may be worse than no messages at all.

Mechanisms and Student-Level Moderators

In this section, we investigate potential mechanisms for the transfer of effects from parents to students. We explore why the effects of view-of-reading content that parents received in text messages only appear to amplify the effects of message personalization for student reading.

Student Self-Reports

Table 1.5 columns (1) and (2) show the self-reported reading behaviors of the students. Column (1) shows that students whose parents received personalized messages reported reading approximately 2.5 percentage points more of their available books on the MORE@Home reading app,[4] but in column (2) there is no effect on visiting the library. There are no significant main effects for either goal setting or view-of-reading in column (1) or (2), but the point estimates are positive for goal setting and the combination view-of-reading conditions, and negative for the

[4] As described previously, all students received access to 6 electronic books and activities via the MORE@Home app, but half also received a set of 10 hardcopy books and additional app activities. Thus, for one half of the sample, the percent of available books read is a proportion of the 16 total books they had access to via the program.

Table 1. 5

Differential Effects of Text Messaging Components on Alternate Student Outcomes

	Student Self-Reports			Qualitative Experience Among App Users			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Proportion of books read (%)	Visited library (%)	Reading motivation	Enjoyed activities	Felt like a good reader	Performance on app activities	Found activities challenging
Panel A - Main Effects							
Personalization vs. Not	0.025*	0.009	0.001	0.023	-0.003	0.083	0.070
	(0.013)	(0.015)	(0.034)	(0.060)	(0.068)	(0.060)	(0.065)
Entertainment vs. Instrumental	-0.019	-0.015	0.005	-0.026	-0.125	-0.095	-0.059
	(0.016)	(0.026)	(0.057)	(0.097)	(0.098)	(0.082)	(0.086)
Combination vs. Instrumental	0.018	0.022	-0.078+	0.079	0.115	0.154+	0.068
	(0.017)	(0.023)	(0.046)	(0.087)	(0.082)	(0.085)	(0.094)
Goals vs. Not	0.002	0.008	0.005	-0.050	-0.031	-0.023	0.078
	(0.010)	(0.014)	(0.029)	(0.061)	(0.068)	(0.074)	(0.080)
Panel B - Two-Way Interaction Effects							
Personalization x Entertainment	-0.001	-0.008	0.039	0.114	0.191*	0.147*	-0.132
	(0.013)	(0.025)	(0.047)	(0.120)	(0.097)	(0.069)	(0.106)
Personalization x Combination	-0.003	0.002	0.002	-0.050	-0.154	-0.093	0.090
	(0.017)	(0.026)	(0.055)	(0.115)	(0.105)	(0.091)	(0.099)
Personalization x Goals	-0.011	-0.005	0.046+	0.058	0.052	-0.026	-0.068
	(0.009)	(0.019)	(0.024)	(0.066)	(0.067)	(0.061)	(0.059)
Entertainment x Goals	-0.031+	0.009	0.102*	0.074	0.054	0.087	-0.024
	(0.016)	(0.025)	(0.047)	(0.085)	(0.095)	(0.079)	(0.092)
Combination x Goals	0.010	-0.053*	-0.125*	-0.134	-0.073	-0.112	0.025
	(0.017)	(0.021)	(0.050)	(0.098)	(0.101)	(0.069)	(0.082)
N	3472	3418	3490	995	995	1155	996

Source: student survey, district administrative records, app-use records

Notes: Point estimates derived from effect-coded regressions that include all treatment factors and their interactions, as well as the following covariates: gender, race/ethnicity, participation in gifted program, participation in Special Education, English learner status, neighborhood SES, language of text messages, baseline reading and math scores, and indicators of the randomization block. Robust standard errors clustered at the family level (in parentheses). +p<0.10, *p<0.05, **p<0.01, ***p<0.001

entertainment-only condition. Looking at the interaction effects for the proportion of books read in column (1), we continue to see that emphasizing the entertainment value tends to mitigate, or counteract, other effects, particularly when combined with goal setting. Students in both the goal setting and the entertainment view-of-reading conditions read 3.1 percentage points fewer of the available books than would have been predicted by the main effects of goal setting and the entertainment-view alone.

We see a different pattern when we consider the differential effects on students' reading motivation as measured by the MMRP in column (3). There are very few main effects, though students receiving a combination of reading values emphasized had fall reading motivations that were lower (ES = -0.08) than students who only received instrumental messages. This negative effect of the combination framing is amplified among students who were also in the goal-setting condition (ES = -0.12). On the other hand, the combination of goal setting and the entertainment value increased student reading motivation (ES = 0.10). We see a similar pattern, though smaller and not significant, when we look at the interaction effects between personalization and the view-of-reading, where entertainment enhances the personalization condition.

Although the strength of these relationships is not consistently statistically significant, these results are particularly interesting because they point in the opposite direction of our main test score results. Students whose parents received the entertainment-only-view messages reported reading less, and if they were also in the personalization or goal-setting conditions tended to score lower on reading assessments. However, the students reported higher levels of reading motivation. This suggests that the entertainment-themed messages did encourage students to see reading as more enjoyable, but that this did not translate into the types of behaviors associated with improving reading skills. In contrast, students who received a

combination of reading values in their message in addition to either goal setting or personalization used the educational app more and scored higher on reading assessments. But these same students reported lower reading motivations when they returned to school in the fall. Additionally, these results suggest that despite not changing observed family behaviors in the app, the framing messages did change the student reading experience outside of the app, possibly through the messages conveyed by parents to their children.

Qualitative App Experience Among App Users

We continue to explore the motivation hypothesis by exploring what students reported about their experiences with the MORE@Home reading app. We know that the app users are a relatively small and unrepresentative subset of our total sample. For example, we know that personalized messages increased app usage, so our app-user sample will be overrepresented by families from that condition. Despite these caveats, students who received personalized messages and those who received a combination of message framing were each more likely to report enjoying the app activities and to find them challenging, though these differences are not statistically significant. There is suggestive evidence that receiving a combination of values increased students' correct answers on the app as well (ES = 0.15), but there are no other significant main effects for the goal setting or reading values conditions.

The interaction effects in Panel B align with the self-reported reading motivation results. Students in entertainment value and personalization conditions in column (5) felt like much better readers (ES = 0.19) than students who were in instrumental value and personalization condition. On the other hand, students in both the combination of values and personalization conditions reported feeling like worse readers than students who were in one or the other (ES = -0.15). These amplification patterns for the values conditions play out in their interaction with

goal setting as well, but the differences are not statistically significant. This may indicate that students who were encouraged to use the reading app for fun felt slightly better about themselves afterwards, whereas students who received targeted "mixed messages" about enjoying and learning from the app tended to feel slightly worse. These exploratory findings reinforce the idea that parental communications to children may have been influenced by message content (the view of reading) in a way that shaped students experience beyond just increasing app usage.

Discussion

In this study, we demonstrate how a factorial design applied to a parent literacy texting intervention can improve both parental engagement with summer reading resources and student academic outcomes. The main contribution of our work is that it clearly demonstrates that variation in the content and tone of messages can lead to important differences in effects. Moreover, our use of a factorial design allows us to estimate not only the impact of each individual component, but also how those components' effects interact.

We find that personalized messages relaying information about student and family reading behaviors improve parental use of an educational reading app across a variety of measures that also translate to positive effects on student reading performance in the fall. Our findings that more personalized messages outperform generic messages on reading comprehension outcomes is consistent with prior work on literacy reminders (Cabell et al., 2019; Doss et al., 2018), but also differs from their experiments in important ways. Our messages include personalized features of the child's name and gender for all students. Our personalization factor instead includes relevant and timely data on the child's performance on the app or book selection for the summer. However, we did not vary the parent's reading activities based upon baseline performance of the child, like Doss et al. (2018). Our effect sizes on reading

comprehension are between these two studies, suggesting that deeper personalization could be an important feature of these messages, particularly for more distal outcomes, like reading comprehension. Another difference that could have led to smaller effect sizes than Doss et al. (2018) is the shorter duration of our texting campaign. Our texting intervention was only 9-weeks during the summer compared with a 10-month period encompassing the school year and the summer. Parents are saturated with texting campaigns and thus, it could be that with deeper personalization over a sustained time-period, allows one to establish themselves as a trusted source of information and rise above the noise.

We hypothesize these information effects could have emerged through two different mechanisms. First, the real-time information about logins and books could be correcting parent's beliefs about how much time they spend on reading activities with their child over the summer. Our information texts caused parents to log-on to the educational app more and our point estimates, while not significant, suggest that parents did more reading activities with their children outside the app as well. Prior work demonstrates that parents tend to underestimate child absences (Robinson et al., 2018; Rogers & Feller, 2018; Smythe-Leistico & Page, 2018). Providing up-to-date information about the number of days a child missed during the school year helps correct these parents' misbeliefs about student absenteeism in early grades. In our context, this suggests that parents may be overestimating how much time they are spending with their child on reading over the summer and the information helps correct these biases. Another helpful component of our interventions could have been that the content of our messages was not focused exclusively on reading activities, but rather included messages related to reading resources, reminders to engage in reading activities, checking-in on progress and in some cases framed reading as both skill building and entertainment. Parents might hit barriers as they try to

implement different strategies and thus giving parents an opportunity to gain confidence in some areas, could help build self-efficacy. Recent work suggests that varying content by domain could be beneficial in sustaining parent's attention on longer campaigns and increasing their self-efficacy in parenting (Doss et al., 2018; Doss et al., 2020).

Our experiment is also the first to rigorously evaluate hypotheses emphasizing particular reading views. Home literacy theory suggests that emphasizing an entertainment-view of reading could lead to improved student motivation and literacy outcomes (Baker et al., 2001). We do not see broad evidence of transfer to self-reported reading motivation or positive feelings about app-based reading activities. Additionally, there are no average effects on student test scores. At the same time, we do find that presenting a combination of the entertainment-view and the instrumental-view amplifies the positive effects of personalization, whereas only presenting the entertainment view detracts from the effects of personalization. Yet, in contrast to the home literacy theory, we find suggestive evidence that for our sample these effects do not operate through student motivation. Students in the personalization and entertainment condition saw statistically significant and positive effects on whether they felt like a good reader, as well as positive effects on a survey of reading motivation, with the converse being true for when both the entertainment and instrumental views-of-reading were present. Thus, these entertainment-themed messages did encourage students to feel more motivated, but this did not translate into the types of behaviors associated with improving reading skills. One potential mechanism could be that adding instrumental messages encouraged parents to practice skill-building with their children, which can be less enjoyable, but would improve their reading scores. Yet, adding some messaging that reading should be fun as well would make this skill building more entertaining than a purely instrumental approach. This type of messages would be particularly useful for

parents if they are also receiving specific feedback about the amount of reading they were doing on the app.

We were surprised by the limited effectiveness of the goal-setting component of our intervention, which tended to have point estimates close to zero across multiple outcomes. Mayer et al. (2018) found large positive effects on the use of an educational app in their comprehensive behavioral intervention (PACT) that leveraged goal setting, reminders, and social pressure. However, there are two key differences that may explain these discrepant findings. First, the goal-setting component of PACT was intensive, involving weekly face-to-face meetings with program staff, which was possible because they were only working with 80 students. Our goal-setting intervention was much lighter touch because it had to be scaled up to reach almost 5,000 families. With so few parents electing to set goals through our online form, the goal-setting factor of our intervention should best be described as a goal-oriented framing of summer reading as opposed to the commitment device used in PACT. In addition, goal setting was just one of three major components of the PACT. We have seen from this study that our own positive effects were largely driven by information-sharing, which aligns more closely with the reminder and social-pressure components of the PACT intervention. It is possible that the large positive effects Mayer et al. find are primarily due to the other components of the intervention.

Policy Implications

The implications of our study are particularly relevant when considered alongside recent syntheses that highlight the difficulty of scaling successful nudging interventions. For example, recent work by DellaVigna & Linos (2020) examined 126 nudge randomized controlled trials in government "nudge units" covering over 23 million individuals. Across their sample, they find statistically significant, but considerably smaller, effects on nudging interventions when

compared to the published literature, which they attribute to publication bias and power (DellaVigna & Linos, 2020). However, the authors do not separately assess content differences, which our work shows could be a crucial feature. As we saw in Figure 1.2, we find that non-personalized messages perform worse for student test scores than sending no text at all, whereas the personalized messages are clearly better. In other successful messaging interventions to increase college enrollment, personalized messages to the students and from the institution staff are standard (Bird et al., 2021; Castleman & Page, 2015; Castleman & Page, 2016; Ideas42, 2015a, Idea42, 2015b; Page et al., 2020). Even though recent work by Bird et al. (2021) tests how these messaging campaigns scale for 800,000 students and find discouraging precisely estimated null effects, they also acknowledge that their ability to personalize the messages was constrained as the intervention scaled. Furthermore, as texting becomes a more ubiquitous form of communication, these results suggest that integrated data-systems allowing for deeper personalization could become increasingly important and valuable tools for changing behavior and student outcomes.

Additionally, in education there has been little causal research on whether changing how interventions are framed for participants could influence behavior and decision-making. For example, all the prior intervention literature on parental texting for younger children focuses exclusively on communicating with parents about building reading skills. Other international research in education nudges has found that reframing the presentation of an unconditional cash transfer as being "for educational purposes" had remarkable effects on parent behavior (Benhassine et al., 2015), but we are aware of no other research in this area. Our work, suggesting that messages can broaden parents' choice sets by exposing them to additional views, could be important for future research not only on parental texting interventions, but for

educational nudges more broadly.

Relatedly, we show that not only does the content of the message matter, but also that it is insufficient to consider how elements of message content like personalization or value-framing operate independently. For our most distal outcome, beginning of third grade exam scores, neither personalization nor an emphasis of particular view-of-reading yields large, statistically significant effects on its own. However, combining these components, personalization with a combination of reading views, is more effective than either one of them alone. Because the message framing seems particularly important for more distal outcomes but not for parental use of the educational app, it suggests that the nuanced differences highlighted in these views of reading could be particularly important for reading behaviors that involve parental engagement in different contexts, such as reading together or talking about reading in the home. The factorial design, which allowed us to show that not all messages are equally effective, can be a valuable tool for other researchers who are hoping to identify the most promising features of an intervention targeting a specific set of desired outcomes in education, health, or civic engagement, or other areas where messages campaigns are popular.

Another key difference between some of the recent attempts to scale messages and successful early literacy interventions is the sustained and consistent contact with parents. One way to maintain this connection is for parents to trust the source of the text messages. In our study, partnering with a school district and individual elementary schools connected our campaign to the existing relationships families have with their school system. Other successful programs have also leveraged these existing relationships. An additional way to sustain the relevance of a messaging campaign is to broaden the scope of the messages. In an age where groups like advertisers and political campaign bombard parents with text messages, a single

reliable source that can vary the types of actions and activities could be beneficial. York et al. (2019) find that their effects of the intervention were much stronger in the second year of implementation, when the messaged touched on a variety of topics, including literacy, mathematics, and socioemotional skills, rather than literacy alone. While they are unable to test the hypothesis directly, it seems likely that the length of the intervention (8 months) benefited from this additional variety and would have otherwise felt repetitive. More recently, Doss et al. (2020) also find evidence of the importance of varying the domain of the content for young students. Similarly, in addition to the app-related resources and reminders, our most successful condition (a combination of reading views in addition to personalized information) included multiple reasons for and ways to engage in reading. At the same time, however, especially if the text can be sustained over longer periods of time, there might be opportunities to apply the lessons learned about message content and framing from text messaging interventions that move beyond "nudges" into sustained assistance and support for parents.

Finally, our work continues to build on the existing literature that a behavioral messaging intervention can improve student outcomes at a much lower cost than other interventions with similar effect sizes. We estimated a back-of-the-envelope cost of this intervention of less than $4 per student, which included the cost to build the connection and integration between a student database and the texting software, the staffing associated with composing each text message, and responding to questions and messages received from families, as well as the cost of sending the text messages through Twilio's platform. While our test score effect sizes are modest, the cost-effectiveness would compare favorably with the Tennessee STAR class size experiment (Schanzenbach, 2006), and other literacy interventions, including Reading Partners (Jacob et al., 2016), Project READS (Kim et al., 2016), and others (Hollands et al., 2013).

Future directions

Our work also offers several lessons for the design of future evaluation of parental texting interventions. First, our text messages were direct to parents, yet only a small proportion of the families responded to our parent survey, which limited our ability to further unpack some of the mechanisms that drove our results. Particularly because factorial design allows researchers to more efficiently explore heterogenous effects, ensuring sufficient effort is dedicated to this data source will be important for future work. Similarly, we solicited our initial goal-setting survey via text message and received a poor response rate on the parental goals, despite being based in goal-setting theory about the importance of plans (Oettingen & Reininger, 2016). Perhaps an alternative in-between our work and Mayer et al., is to call parents for the survey. In other fields, like political science, phone-based interventions encouraging voters to make a plan to vote significantly increased turnout (Nickerson & Rogers, 2010; Rogers et al., 2015) and now is common practice in voter outreach. Third, as noted in other work (York et al., 2019; Kraft and Monti-Nussbaum, 2017) there is learning curve to sending these parental text messages at scale, thus we suggest that district and policymakers commit to these interventions over a period of time in order to reap the benefits after the setup cost. Finally, several recent papers have found smaller or even null effects when text messaging campaigns are scaled to tens or hundreds of thousands of students (Bird et al., 2021; DellaVigna & Linos, 2020), and our own study also struggled to generate parental buy-in when we scaled a goal-setting intervention in a light-touch medium. However, our personalized messages using district-level and app data caused cost-effective improvements in test score outcomes even though our sample was larger than some of the previous literacy-focused studies. While our sample was not on the same scale as Bird et al. (2021), the integration of our data and messaging systems would not face the same constraints of

limited information that they found relying on state- and national-level data to inform their student messages. Thus, further investigation whether scaling using these detailed district-level data to personalize messages can be effective.

Paper 2: Investigating How District-Provided Pre-K Impacts Depend on the Counterfactual

Background

Research has long shown the positive effects of pre-K on student outcomes; the large, positive, and long-term effects of means-tested programs have resulted in a compelling argument that early childhood education is a valuable and cost-effective use of public resources (Heckman, 2006). As a result of this broad evidence base, publicly funded pre-K is a policy that has spread to the majority U.S. States, and the District of Columbia. As of 2018, only 6 states did not provide funding for pre-K (Parker et al., 2018). However, until recently, much of this work focused on federal, means-tested programs, like Head Start, which provide services exclusively to low-income students and families (Deming, 2009; Ludwig & Miller, 2007; US DHHS, 2010).

At the same time, there has been a growing emphasis on the need for universal access public pre-K, such that all students, regardless of family income or "at risk" status, have access to high quality early childhood education. While often called "universal pre-K," these programs do not always guarantee a spot to every interested child. Instead, the universality refers to the eligibility of all students to partake of pre-K services, provided there is sufficient space. Currently, state-level public funding for universal pre-K is still rare: only 8 states and the District of Columbia provide fully or mostly universal pre-K to their constituents (Parker et al., 2018). In addition, some large cities such as Boston and New York have moved toward providing universal programs as well (e.g., Weiland & Yoshikawa, 2013; Rojas et al., 2020). Some research even indicates that universal pre-K may be more effective than targeted programs (Cascio, 2017), and a recent study of the pre-K programs in these eight states finds generally positive effects on student outcomes, particularly in school readiness and math (Barnett et al., 2018). However, rigorous quantitative studies of large, universal programs are surprisingly

sparse.

Much of the current research comes from reports developed by specific universal programs (Phillips et al., 2017). These analyses often present descriptive evidence of changes in outcomes after the implementation of a pre-K program, and others compare pre-K attendees to non-attendees. In general, these studies find positive effects on early academic skills, but typically do not have rigorous causal designs that would account for other differences in pre-K-attending students and their non-pre-K-attending peers (Iowa Department of Education, 2020; Peisner-Feinberg & Schaaf, 2011; West Virginia Department of Education, 2016). A few studies use a regression discontinuity design to compare students who are just old enough to be eligible for pre-K to their peers who are just barely not old enough. These studies typically find large, positive effects on a variety of early academic outcomes in the domains of literacy and math (Gormley & Phillips, 2005; Peisner-Feinberg et al., 2014; Weiland & Yoshikawa, 2013). However, by design, they compare students in pre-K to students who are typically not receiving any sort of educational programming. Unfortunately, they are not able to provide any insight into how these pre-K programs compare to other types of early care options families can access.

Why does it matter what educational services are received by the comparison group? When we are curious about the causal effect of an educational program, we often consider an average treatment effect that compares outcomes between two groups who experience different treatments. While we often think about changing an average treatment effect by changing the outcome of the treated group, a change in the outcome of the control group will have an equal (but opposite) effect on our estimate of the average treatment effect. If everyone in the control group is receiving the same services as the treatment group, there is no treatment-control contrast. In educational research conducted within school settings, our control condition is often

described as “business-as-usual,” or BAU, which means that students in the control group are left to do whatever they would have done in the absence of the intervention being studied. While some concern exists that this BAU condition may not be a sufficiently substantive comparison (Willingham, 2021), it is still useful in uncontrolled educational settings. For example, when evaluating a new math curriculum, the BAU does not expect that the control group teachers will not teach their students math. Instead, it acknowledges that teachers in the control group will be using the most effective strategies that they are currently able to implement without additional supports. In their seminal work, Weiss et al. (2014) argue that the treatment-control contrast is one of three major potential sources of cross-site treatment effect heterogeneity. Their framework shows that variation in the control condition, or the counterfactual, within an individual study can result in variation in the effects of that intervention.

Prior research has shown that understanding this counterfactual has implications for our collective interpretation of the role of early childhood education in supporting student development. For example, in separate studies, Feller et al. (2016) and Kline and Walters (2016) demonstrate that the effects of the federal pre-K program for low-income children, Head Start, differ dramatically based on what the program is being compared to. They document substantial variation in the types of early childhood experiences the control groups receive, including no pre-K, home-based care, and other center-based care. While directly addressing the role of variation in the counterfactual condition in means-tested programs like Head Start is useful, this type of research is even more important when studying universal access pre-K programs. Because universal access programs also serve middle- and upper-income families, students who are not offered places in the public program may be more able to take advantage of potentially high-quality but high-cost pre-K programs that are not accessible to lower income families. A

recent study by Weiland et al. (2020) also considers the distribution of services received by their control group in the context of a universal public pre-K program. While they find that being offered the opportunity to participate in the public pre-K increases the likelihood that study participants will enroll in the school district, this opportunity to participate in public pre-K does not produce significant effects on academic outcomes in elementary school.

Georgia Pre-K

This project investigates the pre-K program in a large metropolitan area school district in the state of Georgia. In the district and throughout Georgia, pre-K is funded and regulated by the Georgia Department of Early Care and Learning Services (DECAL). After beginning in 1992 as a means-tested program to support low-income children, Georgia pre-K transitioned to universal access during the 1995-1996 school year. The program is funded by the state lottery, and in the 2016-2017 school year, this amounted to $350 million, which provides space for approximately 80,000 students across the state. Prior research focused on the Georgia pre-K program includes one study that used aged-based regression discontinuities in eligibility to find positive impacts of the program on a wide range of student literacy and math outcomes (Peisner-Feinberg & Schaaf, 2011). This paper finds that being just old enough to enter the pre-K program improves your reading and math skills a year later. More recent research found that pre-K attendees were more likely to reach reading proficiency in third grade relative to a comparison group (Early et al., 2019), though the research design was not causal.

This current study makes several contributions to research on the Georgia pre-K program and other universal pre-K programs more broadly. First, it considers a specific version of the Georgia pre-K program as administered and managed through a large, diverse school district. Thus, the nature of the treatment is more precisely defined. Second, it draws credible causal

conclusions of the average impact of the program. Using the fact that oversubscribed programs in the district often use lotteries to allocate pre-K spaces, I reconstruct an analytic sample from students who were likely to have participated in one of these lotteries. Finally, it connects existing knowledge on the vital role of the counterfactual in causal research of means-tested pre-K to the growing emphasis on universal programs. Together, this paper helps educational scholars understand that our inferences about the effectiveness of early learning do not just depend on our ability to isolate an internally valid causal impact, but also our ability to understand what exactly we are comparing specific programs and policies to.

District-Run Pre-K

The district's pre-K program is housed within the district's elementary school buildings and is managed by the early childhood department. As of the 2017-2018 school year, the program was available in 45 schools (out of approximately 70 total elementary schools) and served approximately 1,750 students. The district is divided administratively into a set of geographic school communities. Allocation of these programs is dependent on the number of classrooms funded by the state and the ability of the early childhood department to locate physical space on an elementary school's campus where the pre-K classrooms can operate.[5] Additionally, the district has made a concerted effort to consistently offer pre-K programs in low-income communities within the district. Students requiring special education services have access to specifically designated slots in inclusion classrooms that reserve six seats for Special Education students so that they do not need to go through the lottery process. This aspect of the program is funded partially by DECAL, but the district provides and funds additional services beyond what the state provides.

[5] Pre-K classrooms are not allowed to displace K-12 classrooms.

Enrollment in the program is decentralized. In order to apply to the program, families complete a standardized enrollment form and submit it to the school or schools of their choice. If there are more parents interested in enrolling their child in a specific program than there are allocated spaces, schools often conduct lotteries to fairly allocate student spaces. In each school, students are assigned numbers and then numbers are drawn randomly (in a physical, as opposed to electronic lottery). Priority is given to those students who are within the "attendance zone" of that pre-K program.[6] If there are more zoned students than spaces, the lottery will only be relevant for the zoned students. If there are more spaces than zoned applicants, the zoned applicants will receive spaces and the lottery will be conducted among the non-zoned students using the remaining spaces. Because this process is entirely decentralized (organized and managed by each individual school) explicit records of which schools (historically) have used lotteries and which students were in the attendance zone at the time of the lottery were not maintained by the district.

Non-District Pre-K Options

Families living in the district have access to several other early childhood education options, which can be categorized as subsidized and unsubsidized. The subsidized programs include the district-provided program held in elementary schools – this is called public Georgia Pre-K because it is both publicly funded and publicly administered. Other Georgia Pre-K programs are administered by private organizations and companies. Across the state, just over 50% of pre-K classes are privately provided (Georgia Department of Early Care and Learning, 2018). Because of the public funding provided by the state, private GA Pre-K programs are not

[6] For K-12 education, students are geographically assigned (or zoned) for particular elementary schools. While students can theoretically attend a district pre-K program located at any school, they will have to attend their zoned school the following year for kindergarten.

allowed to charge tuition or fees to participating families, making them fully subsidized. A final subsidized source of early childhood education is the state's Head Start program, which served more than 10,000 four-year-old children during the 2016-2017 school year. As the federally funded early childhood option designed to support children from low-income families, Head Start has a robust body of research documenting its characteristics and effects for young students. In addition to these subsidized programs, families also have the opportunity to enroll their children in a variety of private for-profit and non-for-profit pre-K programs. Some of these programs are run by religious organizations. Additionally, some private providers offer both a free Georgia Pre-K program as well as a private, tuition-based pre-K program within the same building.

Drawing on district administrative records, this paper explores the effects of a school-based universal pre-K program in the context of a variety of alternative early childhood educational options. Specifically, I address the following research questions: (1) How can we characterize the use of early childhood education in this district? (2) Among those who express interest in the district program, what is the average impact of attending district-provided pre-K? (3) How are the effects of pre-K attendance related to the availability of counterfactual early learning opportunities?

Methods

Data Sources

All data was drawn from the district's administrative records, of which there are two primary sources. First, the centralized data warehouse contains student demographic information, addresses, enrollment and attendance history, and test score outcomes. The primary outcome I consider is the Georgia Kindergarten Inventory of Developing Skills (GKIDS) Readiness Check.

This assessment is a checklist of basic skills in literacy, mathematics, and development designed to assess the extent to which students are prepared for Kindergarten (Georgia Department of Education, n.d.) and is administered during the first six weeks of the kindergarten year. Scored as a percent correct among the administered items, the GKIDS is entirely formative, in that its purpose is to guide instruction during Kindergarten. Because there are noticeable ceiling effects, I also dichotomize the measure to represent the fact that many students receive perfect scores. This alternate outcome is an indicator of whether the student reached the threshold of 100% completion of the inventory.

Second, in addition to traditional administrative records, the district provided the pre-K waiting lists for each oversubscribed program from the 2007-2008 school year through the 2016-2017 school year. These waiting lists contain the names of all students who ever appeared on the waiting list for each program and their address at the time of enrollment.[7] Pre-K waitlists were matched to district administrative enrollment records with a success rate of 89%, indicating that the vast majority of those interested in attending district pre-K later attended elementary school in the district.

Sample

Descriptive analyses include students who either began pre-K or were eligible for pre-K in the 2007-2008 academic year and kindergarten outcomes are included up through the 2017-2018. Table 2.1 compares the demographics between students who joined the district in pre-K with those who joined in Kindergarten. Across the ten cohorts, almost 30,000 students attended district pre-K, while the remaining 90,000 joined the district in Kindergarten. There are slightly

[7] They sometimes indicate whether students left the waiting list, though this information is provided somewhat inconsistently. Additionally, some students appear on multiple waitlists, or appear both on the waitlist and in a pre-K enrollment file.

more male than female students among both groups, and both groups have sizable populations of Black, Hispanic, and white students, with smaller numbers of Asian students and those who identify as Native American, multi-racial, or other (labeled "Other" in the table). The district pre-K students are significantly more likely to be Black (6 percentage points, $p<0.001$) or Hispanic (11 percentage points, $p<0.001$), to be eligible for free/reduced-price lunch (21 percentage points, $p<0.001$), and to receive English Language services (6 percentage points, $p<0.001$). They are also significantly less likely to be white (-11 percentage points, $p<0.001$) or Asian (-5 percentage points, $p<0.001$). Additionally, pre-K students are four times more likely to be in Special Education than their peers (21% vs 5%), likely as a result of the administrative processes that allocate pre-K spots directly to Special Education students. These statistics represent the

Table 2.1

Demographic Characteristics of Pre-K Attendees to Students Who Joined District in Kindergarten

Characteristic	District Pre-K Attendees	K Joiners	Difference
Male (%)	55.7	51.7	4.0 ***
Race			
African American (%)	44.3	38.2	6.1 ***
Hispanic (%)	24.5	13.5	11.0 ***
White (%)	23.3	34.2	-10.9 ***
Asian (%)	5.0	10.0	-5.0 ***
Other (%)	2.9	4.0	-1.1 ***
English Language Learner (%)	18.0	11.8	6.2 ***
Have Individualized Education Plan (IEP) (%)	21.7	5.2	16.5 ***
FRL-eligible (%)	62.3	41.0	21.3 ***
Number of Students	29,837	88,904	

Notes: *p <0.05, **p<0.01, ***p<0.001

broadest possible set of students in the district, and show that there are large, meaningful differences between the students who first enter the district in pre-K compared with those who enter in kindergarten. A full diagram showing how the other samples, discussed below, were constructed can be found in Appendix B.

Causal Identification Strategy

The demographic discrepancies between the students who join the district in pre-K versus kindergarten (in Table 2.1) highlight the likely selection bias in a simple comparison of pre-K attenders versus kindergarten-joiners, because the characteristics that predict district pre-K attendance tend to also be predictive of lower academic performance, due to systemic inequalities in American society that influence our education system. Additionally, parents who never considered sending their children to district pre-K, either because they not interested in a formal pre-K program, did not know about the district program, or have strong preferences for a different type of pre-K experience, are likely to be different in meaningful ways from the parents who choose that experience; it is also likely that these differences in preferences may reflect other differences in home educational experiences that might predict children's academic outcomes. Thus, to understand the causal effect of attending district-provided pre-K, it is important to identify a more plausible comparison group of students than simply those who joined in kindergarten.

To do this, I consider two such possible analysis samples. First, I limit the comparison group to the students from the district program waitlists. This ensures that all students are coming from families who expressed interest in the district pre-K program. This restriction increases the likelihood that children being compared to each other are similar on other hard-to-measure characteristics reflecting their families interests in and access to early childhood

education. I then also restrict the treatment group to the pre-K attendees from all of the program-years represented in the waitlist control group; this restriction excludes pre-K attendees from programs without a waiting list. The benefit of this sample is that is that it maintains the maximum number of possible students who might have attended the district pre-K program, while also ensuring that the families were expressing interest in the same set of pre-K program locations in each year. This analytic sample is called the Full Waitlist Sample.

At the same time, causal identification is strongest in the presence of a randomized assignment mechanism, which ensures that the decision to attend district pre-K is not associated with the potential benefit that it would provide (Rubin, 1974). To identify a plausibly random process, I leverage the knowledge that the district pre-K program uses lotteries in cases of substantial oversubscription. Working with the district's early childhood department, I used a series of decision rules to identify the specific years in which each center would have used a lottery to allocate spaces in the pre-K program and to identify the students in each of those program years that were likely to have participated in the randomization process. We decided on the following four decision rules:

1. Students with Individualized Education Plans (a designation of receiving special education services) were removed from the sample across all programs and years. Because these students received access to the pre-K program through specially designated slots for students with IEPs, they were considered ineligible for lotteries.

2. Lotteries were limited to programs in years with waitlists of at least 20 students. If a program was undersubscribed or just slightly oversubscribed, program spots were allocated on a first-come first-serve basis as opposed to a lottery.

3. Student zip codes from the waitlist were used to determine the attendance zone of the

waitlist students. The kindergarten elementary school was used to determine the attendance zone of the pre-K attendees.

4. A presence of more than 5 out-of-zone pre-K attendees in a given program-year indicated that all interested in-zone students received slots and that the lottery took place among out-of-zone students. The presence of in-zone students on the waiting list generally indicated that the lottery took place within the attendance zone. However, because students were also often added to the waitlist later (e.g., students who moved to the district over the summer), this was not sufficient to identify whether the waitlist took place among in-zone or out-of-zone students.[8]

Using the first two decision rules, I identified a total of 173 program-year groups in which lotteries were used to allocate spots in the district pre-K program to students, representing 65% of the program-years in the data. Based on the final two decision rules, I identified 6,101 students who were likely to have taken part in those lotteries based on their recorded attendance zones, who comprise the Lottery Based Sample. Among these students, approximately 60% attended the district pre-K, and 40% were found on the waitlist.

Demographic characteristics for these two analytic samples are presented in Table 2.2, using models that include fixed effects for the pre-K program-year, allowing for a clear comparison between students who attended the PreK program at a specific school in a specific year and their respective controls who wanted to but did not attend. The left-hand panel compares the pre-K and waitlist students in the Full Waitlist sample. Despite restricting the comparison to families who desired to participate in the district pre-K program, there are still large, significant differences in the demographic characteristics of the two groups. The district

[8] The addresses of pre-K attendees were based on their most recent address in the district records. Because student mobility within district was high and the zip codes themselves were not a perfect measure attendance zones, the 5-student buffer was created as a compromise.

Table 2.2

Demographic Characteristics and Balance Checks from Quasiexperimental Analysis

	Full Waitlist Sample			Lottery-Based Sample		
Characteristic	**District Pre-K Attendee Mean**	**Control Group Mean**	**Difference**	**District Pre-K Attendee Mean**	**Control Group Mean**	**Difference**
Male (%)	53.4	50.6	2.8 ***	48.4	50.2	-1.8
Race						
African American (%)	50.0	35.7	14.3 ***	28.4	30.4	-2.0 *
Hispanic (%)	24.3	30.8	-6.5 ***	31.4	29.6	1.8
White (%)	17.3	19.4	-2.2 ***	26.7	24.3	2.4 *
Asian (%)	5.9	11.5	-5.7 ***	11.0	12.9	-1.9 *
Other (%)	2.5	2.5	0.0	2.5	2.7	-0.2
English Language Learner (%)	11.2	24.1	-12.9 ***	23.7	20.9	2.8 ***
FRL-eligible (%)	66.1	57.9	8.2 ***	53.0	50.2	2.8 *
Have Individualized Education Plan (IEP) (%)	16.9	5.0	11.9 ***	--	--	
Number of Students	17,410	6,838		3,642	2,459	

Notes: *p <0.05, **p<0.01, ***p<0.001

pre-K attendees were 14 percentage points more likely to be Black, and less likely to be white, Asian, or Hispanic. They were also 8 percentage points more likely to be eligible for free/reduced lunch and 12 percentage points more likely to receive Special Education services. They were, however, 13 percentage points less likely to receive English learner services. The pattern of differences from this sample compared to the overall district demographics reveals that those interested in pre-K are slightly different from those who are not. However, simply restricting the analytic sample to those who were interested is not sufficient to remove observable demographic differences between the pre-K attendees and their counterparts on the waitlist.

The right-hand panel of Table 2.2 shows that the plausible-lottery reconstruction process dramatically reduces the imbalance between our treatment and control groups, suggesting that I have more-or-less recovered the random process allotting pre-K spots. In the Lottery-Based Sample, the pre-K attendees, who were the plausible lottery winners, are still more likely to be Hispanic, white, English-language learners and eligible for free/reduced lunch than the waitlist students from the same lottery, but the magnitude of these differences is very small.

Empirical Model

To estimate the effect of attending district-provided pre-K, we use the same linear model for both the Full Waitlist Sample and the Lottery-Based Sample. Specifically, for student *i* in pre-K program *j* in year *t*:

$$Y_{ijt} = \alpha + \beta PreK_{ijt} + \boldsymbol{X}_{ijt}\boldsymbol{\Gamma} + \omega_{jt} + \epsilon_{ijt}$$

where $Y_{ijt}$is the test score outcomes measured at the start of kindergarten, $PreK_{ijt}$ is the treatment indicator and $\boldsymbol{X}_{ijt}$ represents a vector of student-level covariates including gender, race/ethnicity, ELL status, FRPL status. Finally, ω_{jt} represents a set of fixed effects for the

program-year. In the Full Waitlist Sample, these fixed effects account for average performance differences between pre-K programs in the district as well as any secular time trends in kindergarten readiness. They ensure that our estimates of the effect of attending the district pre-K are drawn from within program-year comparisons of students who attended a specific program at a specific time to their peers who were also interested in that specific program in that year. In the Lottery-Based Sample, these fixed effects also represent the lottery block in which I posit that randomization took place and account for the different treatment-control ratios across those randomization blocks.

Because we do not have the original lottery records that would allow us to know who was initially offered a spot in the pre-K program, we cannot consider analysis from either analytic sample as an intent-to-treat (ITT) analysis. Instead, it is more helpful to think about this estimate as an Average Treatment Effect for a specific subpopulation within the district. By limiting the sample to those who plausibly participated in a pre-K lottery, I am arguing that each of these students (both the pre-K attendees and their counterparts on the waitlist) had a specific probability of receiving a spot in the program. More specifically, we can speculate that within each lottery, all individuals had the same probability, or propensity, for attending the program. Thus, in many ways, this model conceptually aligns with propensity score matching methods (Imbens & Rubin, 2015), where in this case the specific program-year of the lottery perfectly predicts the propensity for treatment. Most importantly, every students' propensity was greater than 0 and less than 1; this means that our analysis also only generalizes to situations in which individuals have probabilistic likelihoods of attending district pre-K, ruling out, for example, the barely oversubscribed programs that assigned spaces on a first-come-first-serve basis.

Results

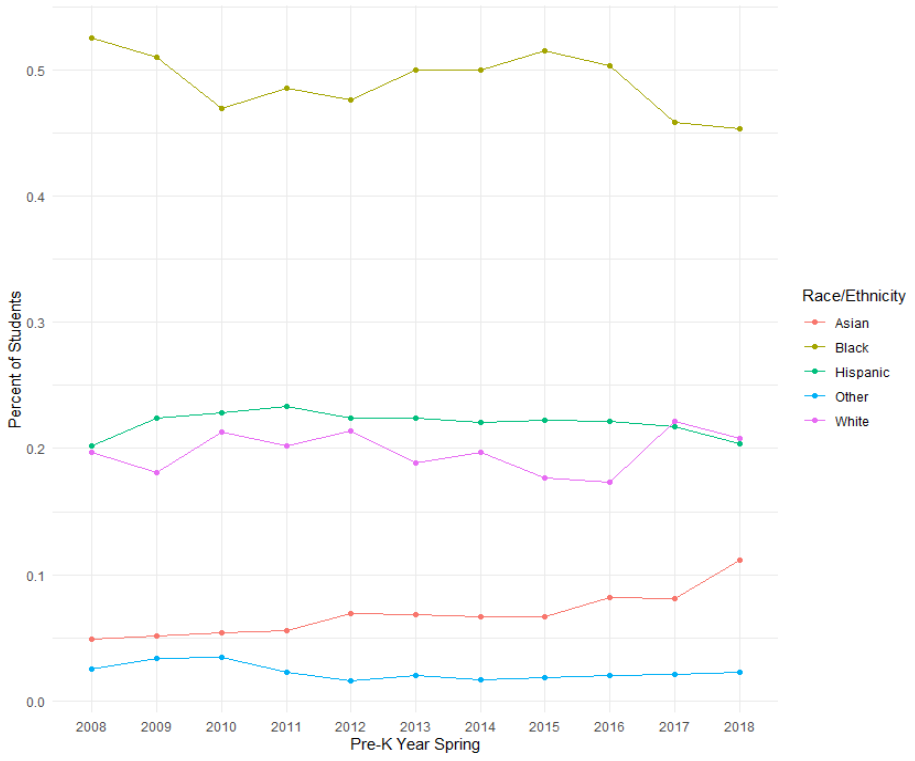
Descriptive Findings

I first consider the demographic characteristics of the students who attend the district run pre-K program. Over the 11 years included in this analysis, the racial distribution of pre-K attendees has largely remained steady. Across all years, the plurality of pre-K students, at least 45 percent, has identified as Black, while white and Hispanic students have typically comprised between 20 and 25 percent of the pre-K population over time (see Figure 2.1). In contrast, the proportion of Asian students has doubled from approximately 5 percent in 2008 to more than 10 percent in the Spring of 2018. Although not displayed in the figure, the percent of pre-K students eligible for free/reduced-price lunch (FRL) has remained steady around 60 percent.

For students in the district studied here, parents are invited to report their child's early childhood education when they complete kindergarten registration. Figure 2.2 shows the reported

Figure 2. 1

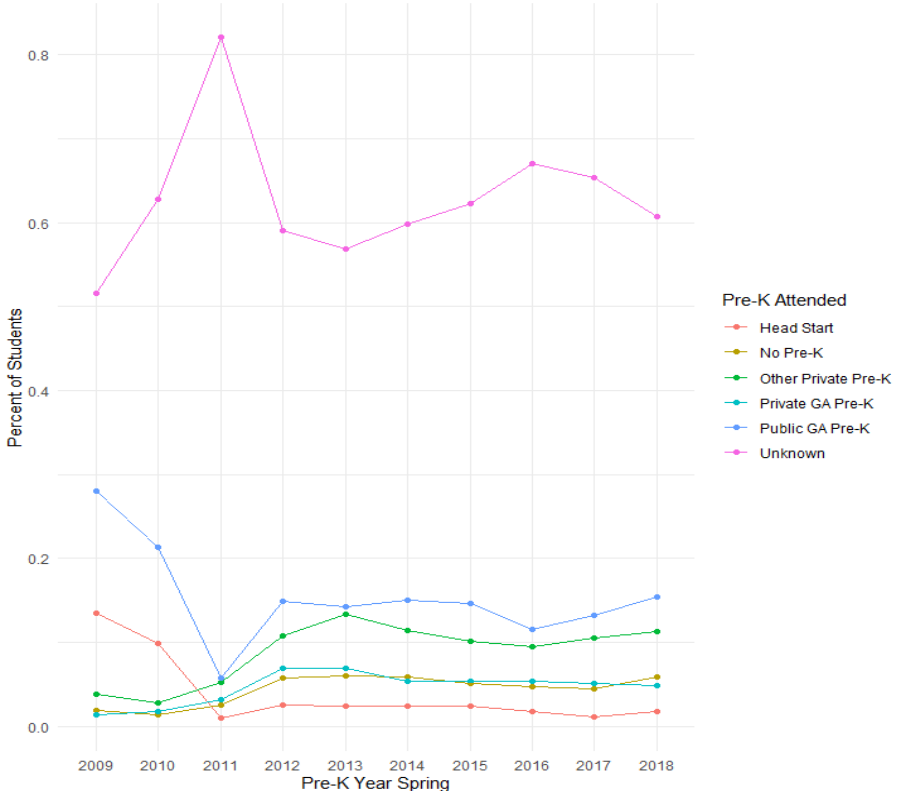
Demographics of District Pre-K Attendees Over Time



values for the kindergarten cohorts during the study years. Consistent with reports by district staff, many parents in each cohort (between 60 and 65 percent) do not complete this field or explicitly report that they do not know the type of pre-K program that their child attended. Pre-K attendance was even less reported in 2011, when this data was unavailable for almost 85% of families. Since 2012, however, a steady pattern as emerged. Over this time, about 1 in 6 students were reported to have attended a publicly provided pre-K program: 15 percent report attending a Georgia Public Pre-K program and an additional 3 percent report having attended a Head Start pre-K program. An additional 10 percent report a publicly funded but privately run Georgia Private Pre-K program, and about 5 percent report not attending pre-K.

Figure 2.2

Control Group Experiences with Pre-K Over Time



While the distribution in usage rates of specific early childhood educational has been stable longitudinally, there is considerable variation across regions in the district, as shown in

Table 2.3. These regions are also very different with regards to socioeconomic status. Only about 20% of students in Communities 1 and 2 are eligible for free or reduced lunch, whereas in Communities 3, 4, and 5, almost 80% of students are eligible for this service. While families in the wealthier communities in the district make use of subsidized pre-K at high rates, between 56 and 59 percent of known enrollment, these communities also have the high rates of enrollment in unsubsidized programs at 36 percent. They also have the lowest rates of known non-attendance at 5 and 8 percent, respectively. Communities 3, 4, and 5, on the other hand, have much higher rates of non-attendance (between 15 and 21 percent). Among families who do report their children attending pre-K, this almost exclusively refers to subsidized programs (either a Georgia pre-K program or Head Start). Fewer than 10 percent of students in each of these communities attend an unsubsidized pre-K program.

Table 2.3

Pre-K experience by School Community Among all Kindergarden Students

	Among Known Pre-K Experience			
School Community	**Percent in Subsidized (N)**	**Percent in Unsubsidized (N)**	**Percent Not Attending (N)**	**Percent Unknown (N)**
Community 1	58.7 (3,954)	36.0 (2,421)	5.3 (358)	55.9 (8,537)
Community 2	55.8 (4,879)	35.6 (3,112)	8.5 (745)	52.3 (9568)
Community 3	69.4 (1,935)	9.3 (258)	21.4 (597)	80.4 (11,443)
Community 4	72.3 (3,580)	8.5 (421)	19.2 (952)	65.1 (9,2440)
Community 5	78.0 (1,927)	6.5 (160)	15.5 (383)	64.8 (4,538)

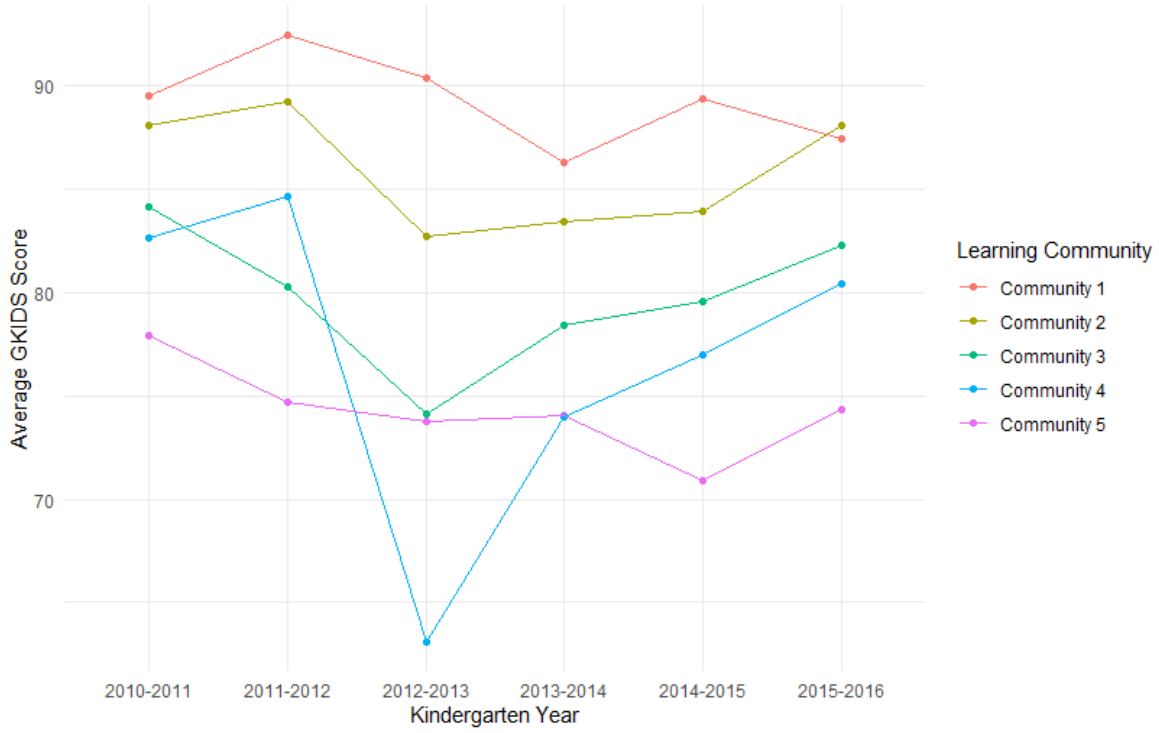
Notes: Community names have been redacted to preserve anonymity. Communities 1 and 2 represent wealthy suburbs. Communities 3 and 4 contain more middle and lower income communities. Community 5 was created to support a historically underperforming set of schools.

While these statistics portray stark differences in pre-K access along socioeconomic lines, it is important to note that overall, the known pre-K experiences represent a small fraction of students. Pre-K data is available at the highest levels for Communities 1 and 2, with about 50 percent of students reporting, and at the lowest levels in Community 3 with only twenty percent of students reporting. Communities 4 and 5 fall somewhere in the middle. Given the low response rates, one concern might be that the families who provide this information are meaningfully different from families who do not. This response bias might carry over and bias any conclusions made about differences across the communities. One way to address this is to consider the demographic characteristics of those with known versus unknown early childhood experiences. A balance check across these two groups, however, shows that there are only minor compositional differences between these groups, with no point estimate being larger than 1.5 percentage points (see Appendix Table B.1). This demographic similarity provides some reassurance that the reports about known pre-K experiences are plausibly representative of the entire district, with the possible exception of the 2011, which from Figure 2.2 seems to be an outlier.

A second way to address this concern about the validity of the self-reported pre-K information would be to consider other measures that provide insight into the quality of early childhood education experiences available to students in in the district. Figure 2.3 shows the incoming kindergarten readiness (GKIDS percent correct) of students in the district, grouped by school community. As might be expected based on the reported pre-K experiences, Communities 1 and 2 consistently have the highest levels of readiness among incoming kindergarten students. On average, students in Communities 3, 4, and 5 are measured as having slightly lower levels of kindergarten readiness, with Community 4 average scores being highly variable over time.

Figure 2.3

Average kindergarten readiness over time, by School Community



All together, these descriptive results suggest that students in the district are exposed to a variety of early childhood educational experiences. Beyond just the district program, thousands of students report attending some form of formal pre-K. And while a large portion of the district does not report their child's early childhood education, these students are not meaningfully different demographically than the students whose families to provide this information. Additionally, we see that there are stark differences across school communities within the district, both in the reported use of early childhood education services and in students' kindergarten readiness when they enter the K-12 system. The school communities with higher rates of free/reduced lunch eligibility are the same communities with lower kindergarten readiness. However, this, and prior research showing larger effects of pre-K among low-income students (Cascio, 2019; van Huizen & Plantenga, 2018) does not sufficiently address why these

differences emerge. The variation in alternative options across communities thus provides a useful tool for categorizing students and exploring differences in the effects of the district program.

Average Treatment Effects from Quasi-Experimental Designs

Because the descriptive results only show broad patterns across all district students, this next section presents causal estimates of attending the district pre-K program relative to the other available early childhood settings. Table 2.4 presents the covariate-adjusted estimates of the impact of attending district-provided pre-K. In the Full Waitlist Sample, district pre-K attendees are compared to all the students on the waitlist for their site. The Lottery-Based Sample is limited to those students who attended or were on the waitlist for a program-year with a lottery and who were likely to have participated in the lottery process. As discussed in the methods section, I consider two separate outcomes drawn from the GKIDS Readiness Checklist assessment, which tests kindergarten students on a series of basic skills. Because the majority of students who complete the assessment are able to complete all skills, thus receiving a perfect score, I explore the impact using both the students' scores as well as indicators of whether they received a perfect score. For the continuous outcome, point estimates of the difference are also provided as effect size, calculated from the overall standard deviation of the outcome. This allows for a comparable estimates across samples within this study and in context of other research.

Among the Full Waitlist analytic sample, I find significant negative effects of attending the district pre-K program. On average, district pre-K attendees score 4.2 percentage points lower on the GKIDS assessment in kindergarten ($p<0.001$). In addition, they are five percentage points less likely than the waitlist students to receive a perfect score. These differences are large

Table 2.4

Effect of Attending District Pre-K on Kindergarten Readiness

Outcome	Sample Size	Point Estimate	Standard Error	Effect Size
Full Waitlist Sample				
GKIDS Percent correct	11,747	-4.19 ***	0.63	-0.15
GKIDS Perfect score (%)	11,747	-5.07 ***	1.10	
Lottery-Based Sample				
GKIDS Percent correct	3649	0.12	0.90	0.00
GKIDS Perfect score (%)	3649	1.01	1.89	

Notes: *p <0.05, **p<0.01, ***p<0.001

relative to the overall performance distribution on this assessment, between 0.1 and 0.15 standard deviations. I find meaningfully different results when I consider the Lottery-Based Sample. Among the students who were likely to have participated in a pre-K lottery, the effect of pre-K on kindergarten readiness is small and not statistically according to both specifications of the outcome. District pre-K attendees scored almost exactly the same on the GKIDS inventory than those who did not attend district pre-K (ES = 0.00, p>0.05). Unsurprisingly, there were just about as likely to receive a perfect score on the assessment. These differences are small in both absolute and relative terms – no larger than 0.02 standard deviations.

The differences in these findings highlight the important distinctions between the two samples. As shown in Table 2.2, there were much larger differences between the pre-K and control groups in the Full Waitlist Sample than the Lottery-Based Sample. While many of the demographic characteristics may be associated with common categories of "advantage" and "disadvantage" and corresponding better or worse academic outcomes, there is not a single clear pattern that would predict how these features might be confounded with treatment effects. However, another key difference is that the pre-K attendees are much more likely to have IEPs,

meaning they need individualized services to support their educational progress. The need for these services is correlated with lower academic performance as well, which may partially explain why these estimated impact of district pre-K is worse in the Full Waitlist Sample. Given these difference, I find the Lottery-Based Sample to provide the more plausibly causal estimate of the impact of the district pre-K program.

Heterogeneous Effects Due to Differing Counterfactuals

There are many cases in which average treatment effects mask substantial heterogeneity, particularly in early childhood programs (Bloom & Weiland, 2015; Feller et al., 2016; McCoy et al., 2016; van Huizen & Plantenga, 2018). Drawing from the literature on the importance of counterfactuals, I use two avenues to explore this. First, I quantify variation based on the administrative region of the pre-K program for both the Full Waitlist and Lottery-Based analytic samples. The top panels of Table 2.5 shows that the large negative effects of district pre-K on kindergarten readiness among the Full Waitlist sample are driven primarily by large negative effects in Communities 1, 2 and 3. In these communities, the waitlist outperforms the district pre-K attendees by at least 0.20 standard deviations ($p<0.001$) While the effects of district pre-K are also negative in Communities 4 and 5, the point estimates are smaller and not significant. The second panel shows that the null average effects among the Lottery-Based sample also masked variation. In Communities 1 and 2, the community-specific effects are similar to the overall effect of district pre-K. On the other hand, the positive effects on school readiness are concentrated among two of the three disadvantaged school communities, with positive effects ranging between 0.1 and 0.2 standard deviations in Communities 3 and 4. Interestingly, there are also large negative effects in Community 5, which was created to provide extra support to a set

Table 2.5

Impact Estimates by School Community

Outcome	Point Estimate	Standard Error	Effect Size
Full Waitlist Sample			
GKIDS Percent Correct			
Community 1	-5.65 ***	1.17	-0.20
Community 2	-12.68 ***	1.63	-0.45
Community 3	-6.62 ***	1.66	-0.24
Community 4	-2.89	1.48	-0.10
Community 5	-3.58	3.25	-0.13
GKIDS Perfect Score			
Community 1	-9.69 ***	2.78	-0.20
Community 2	-16.54 ***	2.81	-0.33
Community 3	-9.71 ***	2.80	-0.20
Community 4	-0.37	2.57	-0.01
Community 5	-8.56	5.54	-0.17
Lottery-Based Sample			
GKIDS Percent Correct			
Community 1	0.74	1.90	0.03
Community 2	-1.40	1.60	-0.06
Community 3	4.66 **	1.50	0.19
Community 4	4.20 *	1.90	0.17
Community 5	-6.24	5.80	-0.25
GKIDS Perfect Score			
Community 1	5.90	4.00	0.12
Community 2	-3.80	3.20	-0.08
Community 3	4.80	3.20	0.10
Community 4	10.60 **	4.00	0.21
Community 5	-25.20 *	12.10	-0.50

Notes: Communities 1 and 2 represent wealthy suburbs. Communities 3 and 4 contain more middle and lower income communities. Community 5 was created to support a historically underperforming set of schools.

*p <0.05, **p<0.01, ***p<0.001

of schools that have struggled for many years with their academic performance.[9]

The internal validity of these estimates relies on the same assumptions as the analysis of the full sample, namely that the sample construction created similar groups. This success with the complete Lottery-Based Sample is now reconsidered within each school community. Table 2.6 presents a series of balance checks across the pre-K and control groups for the Lottery-Based Sample in each school community. In general, the demographic balance is quite good. There are no significant differences in any school community with regards to student gender or free/reduced lunch eligibility. At the same time, there are a few significant differences with regards to ELL status and the general racial composition of the different communities. In Communities 1, 2, 3, and 4, the magnitude of these differences is fairly small. In Community 5, however, there are larger differences, suggesting that the lottery reconstruction process for programs in this community may not have been as successful. As a result, the impact estimates within Community 5 may be less internally valid.

While these results can broadly describe the context in which specific pre-K programs operate, it would be more helpful to know specifically what was happening in the counterfactual. Because the parental reports of pre-K experience had high rates of missingness, I consider the control group outcomes as a measure of the quality of the counterfactual experience. Figure 2.4 shows the average kindergarten readiness (GKIDS percent correct) in each school community for both the Full Waitlist and Lottery-Based analytic samples. In Communities 1 and 2, the collection of counterfactual pre-K options yields good outcomes, particularly for the Lottery-Based Sample. Students in the control group in Communities 3, 4, and 5 have worse

[9] An alternate specification ran a single linear model with a pre-K-by-school-community interaction. The results were substantively the same, and a likelihood ratio test confirmed that the interacted model fit the data better.

Table 2.6

Balance Checks from Lottery-Based Sample, by School Community

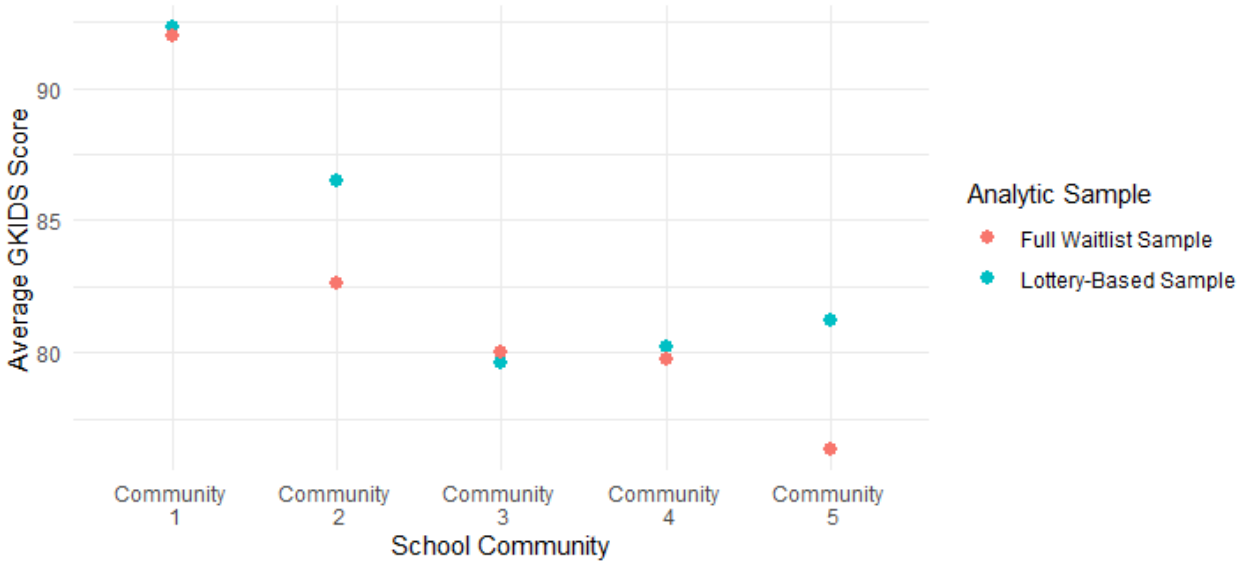
	Community 1		Community 2		Community 3		Community 4		Community 5	
Characteristic	**Control Group Mean**	**Pre-K vs. Control Difference**	**Control Group Mean**	**Pre-K vs. Control Difference**	**Control Group Mean**	**Pre-K vs. Control Difference**	**Control Group Mean**	**Pre-K vs. Control Difference**	**Control Group Mean**	**Pre-K vs. Control Difference**
Male (%)	47.8	0.6	50.9	-3.2	51.7	-0.7	47.1	0.2	52.9	2.2
Race										
African American (%)	10.6	1.5	12.5	-0.7	25.0	-2.5	83.0	-1.5	99.9	-16.9 **
Hispanic (%)	21.5	-3.3	35.8	0.9	48.7	0.4	12.2	4.2 *	0.0	12.8 **
White (%)	45.2	1.8	30.4	1.7	14.8	5.0 **	2.0	-1.2	0.0	0.6
Asian (%)	19.3	-0.8	18.0	-0.2	8.5	-2.8 *	1.2	-0.9	0.0	0.4
Other (%)	3.4	0.9	3.3	-1.7	3.0	0.0	1.5	-0.6	0.0	3.2
English Language Learner (%)	14.9	-6.7 **	29.2	2.1	32.3	5.4 *	5.7	2.3	1.5	7.4
FRL-eligible (%)	25.9	-3.8	42.6	1.5	70.4	1.8	74.1	1.6	86.8	4.5
Number of Students		1,083		1,361		1,900		1,202		466

Notes: *p <0.05, **p<0.01, ***p<0.001

counterfactual outcomes. This reveals that even if the district pre-K attendees performed similarly across school communities, there would still be heterogeneity in the community-specific treatment effects due to the differing effectiveness of the counterfactual.

Figure 2.4

Average kindergarten readiness among the control group, by school community



Discussion

Overall, this study analyzes trends and patterns in access to, as well as outcomes from, a district-run pre-K program in Georgia between 2008 and 2018. I find that program demographics have been largely stable, and that pre-K attendees are substantially different from the students who join the district in kindergarten. In addition to the school-based district pre-K program, students in the district have access to a variety of subsidized and unsubsidized early childhood education options, including other Georgia pre-K programs, private profit and non-profit programs, and Head Start. Early childhood educational experiences are noticeably different across students in the different geographic communities within the district, with higher SES areas having both higher reported rates of unsubsidized pre-K attendance, and higher levels of

kindergarten readiness upon the entry into the K-12 system. Lower SES communities are more likely to report attending a subsidized program or no pre-K program at all and have lower average levels of kindergarten readiness.

Using two potential analytic samples based on the district's program waitlists, I also estimate the average impact of attending the district pre-K program. I find evidence of significant negative effects on kindergarten readiness on the magnitude of 0.10-0.15 standard deviations among the broader Full Waitlist Sample ($p<0.001$) and small and nonsignificant effects no larger than 0.02 standard deviations among the Lottery-Based Sample. Both analytic samples address a common concern in matching-based quasi-experimental research: while it is possible to match on observed characteristics, it is harder to believe that a researcher has matched on unobserved characteristics as well. By restricting the analysis to families who have expressed interest in attending specific district-run pre-K programs, these samples are able to capture one typically hard-to-measure feature affecting students' early academic outcomes – a family's interest in and commitment to early childhood education. However, balance checks among the two samples indicate that relying on the waitlist is not sufficient to identify similar families. The Full Waitlist Sample control group is sufficiently different from the district pre-K attendees that it suggests the impact estimates will suffer from significant selection bias. The improved demographic balance in Lottery-Based Sample supports the claim that the lottery reconstruction process was able identify a set of program-years and a subset of the treatment and control groups where a randomized allocation of district pre-K spots took place. This combination of matching on the typically unobserved characteristic of interest in early childhood education and demographic characteristics makes the impact estimates from this sample more plausible.

There are several possible explanations for the lack of balance in the Full Waitlist

Sample. In many of the district's programs, the first-come first-serve enrollment process will mean that families on the waitlist are all those who signed up later. Recent research from the New York City kindergarten enrollment system finds that some groups of families, including those with English Language learners and those experiencing housing instability systematically join the enrollment process later than students from families not facing these systemic barriers (Condliffe & Balu, 2019). Similar gaps in access to program and school information across racial and socioeconomic lines have also been documented in other urban areas across the country (Condliffe et al., 2015; Owens et al., 2016; Shapiro et al., 2019).

The difference between the two analyses may also reflect differences in the program-years included in the samples. The Full Waitlist Sample includes programs that are barely over-subscribed. To the extent that families are aware of meaningful differences between programs and are choosing based on perceived quality, it may be possible that this sample includes a broader range of program effectiveness than the Lottery-Based Sample. The opposite concern, that lottery estimates relying on large amounts of oversubscription may overestimate average program effects, has been raised in other studies using that design (e.g., Bloom & Unterman, 2014; Weiland et al., 2020). Overall, however, these pre-K programs represent a more standardized form of treatment than many of the other contexts in which lottery reconstructions have been used to estimate causal effects. Due to curricular requirements tied to the public funding as well as the centralization of the program at the district level, I am less concerned about treatment-based differences influencing the causal estimates in this study.

A major contribution of this paper is to connect the impacts of attending the district pre-K program to the counterfactual experiences of the different school communities. Across both samples, the district program fares worse in the higher SES communities and better in the lower

SES communities; this is consistent with other recent research finding that universal pre-K programs have particularly large effects for low-income students (van Huizen & Plantenga, 2018). However, other research on universal pre-K does not address why its effects may be likely to be larger for low-income students. A recent study by Pearman (2020) explores potential reasons that may explain differences in pre-K effects across communities with different levels of poverty, though it is based on the Tennessee pre-K program which only served low-income students. Even among low-income students, however, Pearman finds that the positive effects on reading are larger in high-poverty communities and hypothesizes that this may be due to a combination of "risk factors" like living in a single-parent household, or a lack of alternate care providers in the neighborhood. However, he is unable to connect the alternate settings to the control group in the Tennessee context.

Other research explores the role of the counterfactual experience more explicitly. Feller et al. (2016) and Kline and Walters (2016) find that Head Start impacts are predominantly concentrated among students who otherwise would have been in home-based care, with minimal effects for students who would have otherwise attended a different care center. Weiland et al. (2020) are also able to identify the specific counterfactual choices of their control group in the context of a universal pre-K program, but do not look at how these differences may have been related to their observed effects. This paper contributes to this conversation by connecting the pieces between these different studies. Like Pearman (2020), I use community to define the relevant unit of analysis for identifying variation in treatment effects. However, instead of relying on broad measures of program density, I use administrative data to identify the distribution of pre-K experiences within my sample for each community. These communities matter not because they contain more or fewer students of a particular SES, but because families

in these communities use alternate care providers at different rates, which result in different control group outcomes. Using the control groups kindergarten readiness as a proxy for the quality of the counterfactual educational quality, I show that communities with larger effects tend to have the lowest levels of kindergarten readiness among their control groups – in both the Full Waitlist and the Lottery-Based analytic samples.

Taken together, these results contribute to the growing literature on the role of counterfactual conditions in understanding the effectiveness of early childhood education. More specifically, while community poverty levels may be an easy-to-capture moderator of effects, this is likely due to the interaction of poverty and access to alternative high-quality early childhood education. Unfortunately, this study is unable to provide much information about the structural or curriculum components of the counterfactual options that might make them individually more or less effective; instead, its conclusions rely on observing the output of these programs in terms of students' readiness for kindergarten. Looking to the future, this information would be helpful for extending research on counterfactual quality, particularly given new research revealing that continued access to high quality instruction and curricular alignment across grades helps sustain the effects of pre-K programs (McCormick et al., 2019; Unterman & Weiland, 2020).

Paper 3: Inference in Sequential Multiple Assignment Randomized Trials

Background

Adaptive Interventions

Adaptive interventions are interventions that begin with an initial implementation but may be later modified to address the perceived responsiveness and needs of individual children. In elementary literacy, for example, the Response to Intervention (RtI) framework can be seen as a canonical example of an adaptive intervention designed to ensure all children receive the supports they need to be successful readers (Balu et al., 2016). Under RtI, all students receive Tier 1 literacy instruction; students who are not reading at the desired level are flagged to receive Tier 2 services, or more intensive supports for reading development. If a student is still not improving their reading skills with Tier 2 services, they can be escalated to receive Tier 3, or the most intensive, services, which often include one-on-one interventions from a highly qualified provider. Adaptive interventions are thus a type of personalized intervention designed to provide individual students with the specific resources needed to address their skill gaps. Personalized interventions are prevalent in literacy (e.g., Connor & Morrison, 2016; Jones et al., 2016; Joseph 2018), and there is some evidence that specific examples of the personalized intervention approach can be effective at preventing the need for later interventions (Connor, 2017).

At the same time, promising examples of personalized or adaptive interventions that have been effective in small experiments often fail to replicate when scaled-up (Balu et al., 2016; Denton et al., 2010). This may be in part because in many of these interventions, a practitioner must correctly diagnose student difficulties, identify appropriate intervention strategies to support those difficulties, and implement the selected strategies. This multi-step process is complicated, and an individual program or system could fail at any one of these three steps. This

complexity means that failure to find positive effects could be due to a failure of diagnosis or of targeting the supports as opposed to a failure of a personalized intervention more broadly. To maximize the efficacy of an adaptive intervention, researchers must address these different components during the process of developing and refining the adaptive intervention. Notably, standard experimental designs commonly used to assess the effects of educational interventions – such as two-arm RCTs or comparison groups – are not well-suited to statistically distinguish or identify where these interventions may go awry. In the space of education, these mechanisms have historically been addressed through theory and qualitative evaluations of intervention implementation at large. As adaptive interventions growing in popularity and scale due to the increasing integration of technology into educational settings, it is important for the field to consider more rigorous statistical approaches to develop, test, and refine adaptive interventions.

Sequential Multiple Assignment Randomized Trials

Sequential Multiple Assignment Randomized Trials (SMARTs) are designed to assist in the development, refinement, and testing of adaptive interventions. In SMART trials, individuals are initially assigned to a treatment condition and their responsiveness to this initial condition is measured at a set point in time using what is called a "tailoring variable," often an intermediate outcome. Individuals who achieve a pre-determined level of the tailoring variable are designated as "responders," whereas those who do not achieve this level are designated "non-responders." Based on this responder status, individuals then receive the second phase of treatment. Individuals who are identified as responders to the initial intervention go on to receive one set of later treatments (e.g., a continuation of the services they were receiving in Phase 1), while non-responders are typically randomly assigned to receive different types of additional supports or modifications to improve their outcomes (Collins et al., 2007).

One advantage of the SMART approach is that the specific design can be tailored to a variety of different circumstances, while still supporting the overarching goal of developing and refining an adaptive intervention. One of the most common features across all SMART designs is that Phase 1 includes two active treatments which are being compared to one another, as opposed to one treatment being compared to a control condition. This allows researchers to address "which way (Y or Z) is better to address X problem" instead of simply "does Y solution address X problem." But often, the specific components of Phase 2 and how they are assigned differ across studies. Nahum-Shani et al. (2012) provide several different examples, with accompanying figures, of how the specific designs might be adapted to the researcher's needs – we highlight just a few here. For example, while many SMARTs randomize responders to receive different supports in Phase 2, some SMART studies just continue providing responders with their Phase 1 treatment even once they enter Phase 2; this may be a particularly desirable research design if there are severe resource constraints and the success of the initial treatment also bodes well for long term outcomes. Another common variation in SMART designs is that the secondary support options in Phase 2 may be defined in reference to the initial treatment from Phase 1– one option for non-responders is to receive either an augmented version of their initial treatment or a different support added onto their initial treatment. This inherently flexibility in the SMART framework – including the abilities to decide who is randomized in Phase 2 and to define the Phase 2 supports differently across Phase 1 treatments – makes it appealing to researchers. And while this design has primarily been used in psychological and medical settings (Almirall et al., 2014; Kidwell & Hyde, 2016; Pelham et al. 2016), it has recently been introduced for conceptually for educational interventions (Almirall et al., 2018)

and applied to issues such as chronic absenteeism and elementary literacy (Heppen et al., 2020; Kim et al., 2019).

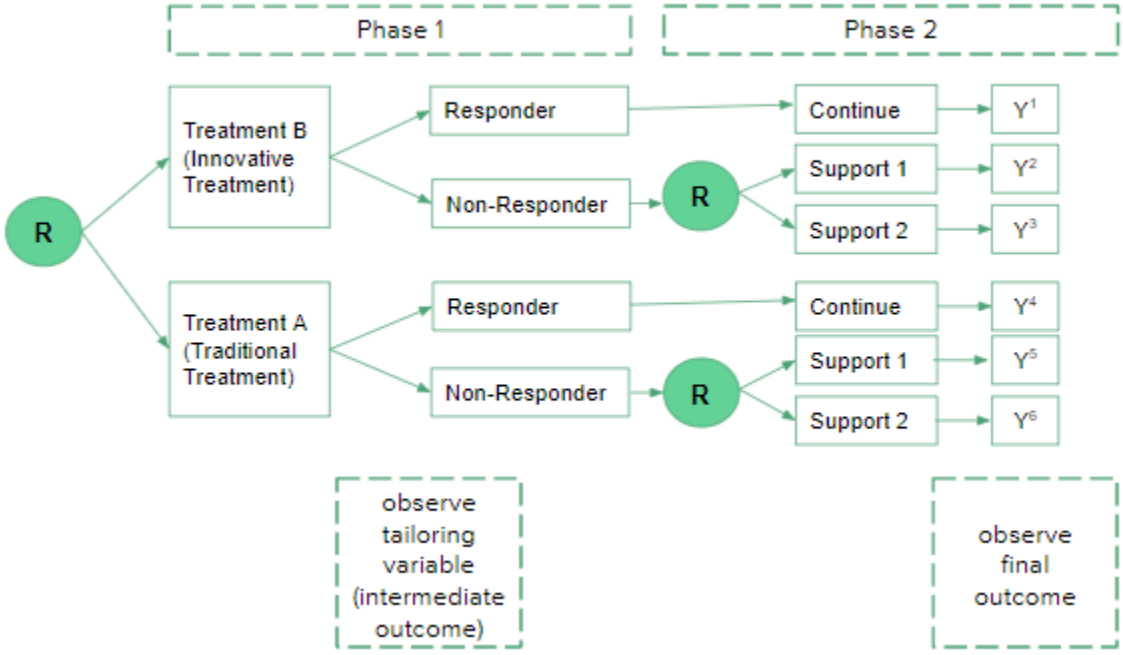
Because there are multiple waves of random assignment, SMART designs of all types offer researchers the opportunity to causally estimate a variety of treatment effects with confidence in their internal validity. Figure 3.1 presents a prototypical SMART design, adapted from Almirall et al. (2014), where in Phase 1 students are initially assigned to one of two treatments: either Treatment A, a traditional intervention, or Treatment B, an innovative intervention that is being compared to Treatment A. Based on the tailoring variable, some students are identified as responders, because they behave in the desired way following the initial treatment. Those identified as responders at the end of Phase 1 continue to receive their initial treatment in Phase 2, whereas non-responders are randomly assigned to one of two additional treatments, which are designated in the figure as Support 1 or Support 2. This experimental design results in 6 final groups within the sample (with observed outcomes Y) that are defined by their realized treatment trajectories. Despite this elegant design with randomization to treatment conditions, causal estimands are not well-defined in the existing SMART literature. Nahum-Shani et al. (2012) do offer sample-based estimators of the "main effects" of Phase 1 and Phase 2 treatments-contrasts based on the six subgroups defined by the design.[10] Below, we present these commonly used estimators, which we argue are not sufficiently paired to causal estimands of interest.

First, the main effect for comparing the two initial interventions can be simply defined as

[10] SMART designs are also interested in comparing the different embedded adaptive treatment regimes to one another and also explore covariate-by-treatment interactions which they call "optimizing" analyses (for a description, see Almirall et al., 2014). Though beyond the scope of this paper, these comparisons will likely face many of the same issues as the main effect estimators because they also rely on standard regression estimators.

Figure 3.1

Example SMART Design with Re-Randomization for Non-Responders



the difference in pooled outcome of the subgroups associated with each initial intervention:

$$\widehat{ATE}_{Phase1} = \frac{1}{n_1+n_2+n_3}\sum_n(Y^1 + Y^2 + Y^3) - \frac{1}{n_4+n_5+n_6}\sum_n(Y^4 + Y^5 + Y^6) \quad \text{Eq. (1)}$$

where the superscripts refer to the different sample subgroups in Figure 3.1. So, for example, the relative average treatment effect of starting the SMART with the innovative Treatment B as opposed to the traditional Treatment A is just the average difference in outcomes between all subgroups who received Treatment B initially (Y^1, Y^2, and Y^3) and all subgroups who received Treatment A initially (Y^4, Y^5, Y^6).

Second, researchers are often interested in the main effect between the second-phase options. The specific types of Phase 2 supports can, of course, differ across interventions. For example, in some cases, Support 1 might involve increasing the dosage of the first-phase intervention and Support 2 might be augmenting the first-phase intervention with the alternative

first-phase intervention. In other cases, the supports might be entirely different approaches, or two versions of a supplemental intervention that was not administered in the first phase. Regardless of what specifically the two options are, the main effect compares them to each other. In this case, the authors recommend a similar approach of pooling across groups that received the same second-phase treatment:

$$\widehat{ATE}_{Phase2} = \frac{1}{n_2+n_5}\sum_n(Y^2 + Y^5) - \frac{1}{n_3+n_6}\sum_n(Y^3 + Y^6) \qquad \text{Eq. (2)}$$

If the second phase of random assignment provides a 50% treatment allocation to each of the two supports, this estimator is simply the weighted average of the two effects of the second phase estimated within blocks of the first-phase assignment, where the weights are equal to the proportion of non-responders found in each of the first-phase treatment arms. We show this below by rearranging Equation (2) and strategically replacing subsample size notation:

$$\widehat{ATE_{NRr}} = \frac{n_2\overline{Y_i^2}}{n_2+n_5} + \frac{n_5\overline{Y_i^5}}{n_2+n_5} - \frac{n_3\overline{Y_i^3}}{n_3+n_6} - \frac{n_6\overline{Y_i^6}}{n_3+n_6}$$

$$\widehat{ATE_{NR}} = \frac{n_2\overline{Y_i^2}}{n_2+n_5} - \frac{n_3\overline{Y_i^3}}{n_3+n_6} + \frac{n_5\overline{Y_i^5}}{n_2+n_5} - \frac{n_6\overline{Y_i^6}}{n_3+n_6}$$

$$\widehat{ATE_{NR}} = \frac{n_2\overline{Y_i^2}}{n_2+n_5} - \frac{n_2\overline{Y_i^3}}{n_2+n_5} + \frac{n_5\overline{Y_i^5}}{n_2+n_5} - \frac{n_5\overline{Y_i^6}}{n_2+n_5}$$

$$\widehat{ATE_{NR}} = \frac{n_2}{n_2+n_5}(\overline{Y_i^2} - \overline{Y_i^3}) + \frac{n_5}{n_2+n_5}(\overline{Y_i^5} - \overline{Y_i^6})$$

$$\widehat{ATE_{NR}} = \frac{2n_2}{2(n_2+n_5)}(\overline{Y_i^2} - \overline{Y_i^3}) + \frac{2n_5}{2(n_2+n_5)}(\overline{Y_i^5} - \overline{Y_i^6})$$

$$\widehat{ATE_{NR}} = \frac{n_2+n_3}{(n_2+n_3+n_5+n_6)}(\overline{Y_i^2} - \overline{Y_i^3}) + \frac{n_5+n_6}{(n_2+n_3+n_5+n_6)}(\overline{Y_i^5} - \overline{Y_i^6})$$

In relying on a weighted average of these two separate estimates, we hypothesize that existing estimators are making a latent assumption about the comparability of the observed non-

responders from each of the first-phase treatments. A less restrictive version of this assumption is simply that individuals' responsiveness to the first phase is uncorrelated with the differential effects of the second phase supports. Given that the entire premise of SMART designs acknowledges that different individuals respond differently to treatment options, why would we expect the individual treatment effects for Phase 1 and Phase 2 to be uncorrelated? This paper seeks to clarify inference in SMART designs in the context of these different assumptions. First, we use principal stratification to identify causal estimands in SMART designs. We then consider the extent to which standard analytic techniques like those proposed by Nahum-Shani et al. (2012) allow us to recover those estimands.

Using simulations, this paper explores when differences between the non-responder groups can invalidate traditional inference from SMART designs. Specifically, using the potential outcomes framework (Rubin, 1974), we explore the validity of existing estimation methods for understanding effects in SMART designs by asking the following questions:

1) To what extent can standard estimators recover well-defined causal estimands of Phase 1 and Phase 2 impacts under different models of responsiveness?

2) How does the validity of these estimates change when there is a correlation between responsiveness to the first stage of intervention and the individual effects of the Phase 2 supports?

It might seem odd to focus on traditional causal inference techniques when the purpose of a SMART is not to test the efficacy of an intervention relative to an untreated control group. However, the stated purpose of SMARTs is to optimize the development of an adaptive intervention in the social, behavioral, and educational sciences, based on the outcomes this intervention provides to participants. In order to make these type of prescriptive comparisons

about relative effectiveness, researcher must draw causal conclusions. And do to that well, they must use rigorous causal inference.

Modeling Responsiveness Using Potential Outcomes

Responsiveness Defines Principal Strata

To successfully model treatment effect estimates, we must first define individual outcomes under each of the different hypothetical treatment regimes. These potential outcomes (Rubin, 1974) form the foundation for identifying the average treatment effect estimand and appropriate estimators. Beginning with the first phase of the prototypical example in Figure 3.1, individuals can either be assigned to Treatment A, which we consider the traditional treatment, or Treatment B, the innovative treatment. At this point, individuals may respond differently to their assigned treatment – some will "respond" and see improved outcomes, whereas others will be "non-responders." While responsiveness is an endogenous individual characteristic that is only observed after, and in response to, treatment, we can identify and define endogenous groups of individuals based on how they would hypothetically respond to the different treatment options. This approach is called principal stratification, as each endogenous subgroup is defined as a unique stratum of the sample (Frangakis & Rubin, 2002). Table 3.1 defines the 4 principal strata inherent in Phase 1 of a SMART design like the one in Figure 3.1. Those individuals with poor outcomes under both treatment conditions fall into Stratum 1: Never responders. Those who do well under both treatments comprise Stratum4: Always responders. Those who only do well under (the innovative) Treatment B but do not respond to (the traditional) Treatment A comprise Stratum 2: Innovative responders. Finally, those who only respond to (the traditional) Treatment A but do not benefit from (the innovative) Treatment B comprise Stratum 3: Traditional responders.

Table 3.1 also shows how the individual's potential responsiveness under each initial treatment is related to their intermediate tailoring variable and stratum membership. As with all causal studies, the potential outcomes depend on the research design. While a traditional, two-arm RCT involves a single assignment mechanism ($Y_i(z), z \in \{0,1\}$), the multiple assignment mechanisms in a SMART make defining potential outcomes more complex. To begin with the simplest case, we first consider the potential outcomes for the responsiveness $R_i(z)$, with $z \in \{A, B\}$. Table 3.1 also shows how the principal strata provide bounds on values of each potential outcome relative to c, the cutoff value used to identify responders and non-responders. For Never Responders, both $R_i(A)$ and $R_i(B)$ fall below the responder threshold, whereas for Always Responders, both $R_i(A)$ and $R_i(B)$ fall above the responder threshold. Innovative Responders and Traditional Responders both have one potential outcome on each side of the threshold, though the positions are reversed between the two groups.

Table 3.1

Principal Strata and Potential Responsiveness in a SMART

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="2">Responsiveness under Treatment B

(Innovative Treatment)</td></tr>
<tr><td>Low/unimproved intermediate tailoring variable</td><td>High/improved intermediate tailoring variable</td></tr>
<tr><td rowspan="2">Responsiveness under Treatment A

(Traditional Treatment)</td><td>Low/unimproved intermediate tailoring variable</td><td>1. Never responders

$(R_i(A), R_i(B)) < c$</td><td>2. Innovative responders

$R_i(A) < c \leq R_i(B)$</td></tr>
<tr><td>High/improved intermediate tailoring variable</td><td>3. Traditional responders

$R_i(B) < c \leq R_i(A)$</td><td>4. Always responders

$(R_i(A), R_i(B)) \geq c$</td></tr>
</table>

Applying Potential Outcomes to Phase 2

After observing the intermediate tailoring variable, responders continue receiving their Phase 1 Treatment, and non-responders are re-randomized into their Phase 2 supports. This results in a more complex set of potential outcomes for the final outcome, Y_i. Table 3.2 shows the potential outcomes for each principal stratum as a function of both the Phase 1 and Phase 2 treatment assignments: $Y_i(z, w)$, where, as before, $z \in \{A, B\}$ and now, $w \in \{1,2,-\}$ reflecting that individuals can receive either Support 1, Support 2, or, if they responded to their initial treatment, the continued provision of their initial treatment. In the prototypical SMART design we consider, there are different potential outcomes for the different strata. Different instantiations of the SMART framework (for example, those that also re-randomize responders to receive additional supports) may have different sets of potential outcomes based on the specifics of randomization within the specific SMART design.

Table 3.2

Potential Outcomes under Example SMART design with Re-Assignment for Non-Responders

	Non-responder to Treatment B	Responder to Treatment B
Non-responder to Treatment A	1. Never responders $Y_i(A,1)\ \ Y_i(A,2)$ $Y_i(B,1)\ \ Y_i(B,2)$	2. Innovative responders $Y_i(A,1)\ \ Y_i(A,2)$ $Y_i(B,-)$
Responder to Treatment A	3. Traditional responders $Y_i(A,-)$ $Y_i(B,1)\ \ Y_i(B,2)$	4. Always responders $Y(A,-)$ $Y_i(B,-)$

With the potential outcomes now well-defined, it is possible to define estimands of interest – for example, an average treatment effect (ATE), or a Principle Causal Effect (PCE), which can be thought of as an average treatment within a particular principal stratum. We define the Principle Causal Effects for each stratum $S \in \{NR, IR, TR, AR\}$ as follows. For an individual in the Never Responder stratum, the individual treatment effect (ITE) of receiving initial Treatment B as opposed to initial Treatment A in Phase 1 is:

$$ITE_{1i,NR} = E_w[Y_i(B, w) - Y_i(A, w)]$$

where for each initial treatment, you take the expectation over the possible Phase 2 supports. In practice, with 50/50 assignment ratios, this amounts to the average across each Phase 2 condition (i.e., $ITE_{1i,NR} = \frac{Y_i(B,1)+Y_i(B,2)}{2} - \frac{Y_i(A,1)+Y_i(A,2)}{2}$), but with other assignment conditions it would be weighted average. For these Never Responders, we define the individual treatment effect of receiving Support 2 as opposed to Support 1 in Phase 2 as:

$$ITE_{2i,NR} = E_z[Y_i(z, 2) - Y_i(z, 1)]$$

where now you take the expectation over the initial treatment conditions. Once you have the individual treatment effects, the Principle Causal Effect is simply the expectation of those ITEs over the stratum:

$$PCE_{1,NR} = E[ITE_{1i,NR}]$$

$$PCE_{2,NR} = E[ITE_{2i,NR}]$$

In some ways, the Never Responder stratum is the easiest to define because all participants are re-randomized into Phase 2 supports. The other extreme is the Always Responder stratum, where no-one receives a Phase 2 support. In this case, the $ITE_{2i,AR}$ is undefined, but the first-phase effects can be written:

$$ITE_{1i,AR} = Y_i(B, -) - Y_i(A, -)$$

$$PCE_{1,AR} = E[ITE_{1i,AR}]$$

For the Traditional Responder and Innovative Responder strata, the individual causal effects and principle causal effects are defined similarly:

$$ITE_{1i,TR} = E_w[Y_i(B,w)] - Y_i(A,-)$$

$$ITE_{2i,TR} = Y_i(B,2) - Y_i(B,1)$$

$$PCE_{1,TR} = E[ITE_{1i,TR}]$$

$$PCE_{2,TR} = E[ITE_{2i,TR}]$$

In the Traditional Responder stratum, there is only one potential outcome for individuals who receive initial Treatment A. Thus, the Phase 1 ITE only requires aggregating across potential outcomes $Y_i(B,w)$, and the Phase 2 ITE is a simple difference in means that is only defined for Phase 1 Treatment B. Innovative Responders follow the same pattern, only swapped:

$$ITE_{1i,IR} = Y_i(B,-) - E_w[Y_i(A,w)]$$

$$ITE_{2i,IR} = Y_i(A,2) - Y_i(A,1)$$

$$PCE_{1,IR} = E[ITE_{1i,IR}]$$

$$PCE_{2,IR} = E[ITE_{2i,IR}]$$

Ultimately, however, intervention researchers are often thinking about estimating the average treatment effect across the entire population for the first and second phases (as shown in Equations 1 and 2), as opposed to the principal causal effects for each stratum. This Average Treatment Effect (ATE) can be defined as a weighted average of these principal causal effects:

$$ATE_1 = \pi_{NR}PCE_{1,NR} + \pi_{AR}PCE_{1,AR} + \pi_{TR}PCE_{1,TR} + \pi_{IR}PCE_{1,IR}$$

$$ATE_2 = \frac{\pi_{NR}PCE_{1,NR} + \pi_{TR}PCE_{1,TR} + \pi_{IR}PCE_{1,IR}}{\pi_{NR} + \pi_{TR} + \pi_{IR}}$$

where the weights π_s represent the proportion of the population that falls into stratum S. Because the Phase 2 ITE and PCE is undefined for the Always Responders, the Phase 2 ATE must be rescaled by the proportion of the population in the remaining three strata and will only be generalizable to those strata.

Connecting Causal Estimands to Estimators Used in Current Practice

Using these principal strata, we can then define the different estimators provided by Nahum-Shani et al. (2012) in terms of the population parameters they estimate. While Equations 1 and 2 represent the sample-moment-based estimators, standard analytic practice involves typical OLS regression involving indicator variables to distinguish between the relevant groups. To compare these estimators, we must first demonstrate which principal strata appear in the six observed subgroups from our design. Subgroup 1 will contain Traditional Responders and Always Responders; subgroups 2 and 3 will contain Never Responders and Innovative Responders. Subgroup 4 will contain Innovative Responders and Always Responders, and subgroups 5 and 6 will contain Never Responders and Traditional Responders. We can represent these subpopulation effects as weighted averages of the principal causal effects that are well-defined above. However, we cannot do the opposite: it is impossible to calculate the principal causal effects using the subpopulation outcomes, because we do not know the relative presence of two strata within each subpopulation.

This notation also highlights the underlying assumption of the existing estimators: by defining responder status only based on the observed tailoring variable, they assume away the different stratum and assume that all responders are drawn from the same population. Thus, we simply have an amorphous group of "responders," instead of the Traditional, Innovative, and Always Responders as defined by the principal strata. We call this assumption the Common

Population Assumption. More specifically, when the Common Population Assumption (CPA) holds, responsiveness is the same under both first stage interventions: $R_i(A) = R_i(B)$. As a result, the Traditional Responder and Innovative Responder subgroups do not exist, and all participants are either an Always Responder or a Never Responder.

Simulations

General Framework

The simulation study investigates the validity of inference in SMART designs by examining the extent to which standard analytic approaches (namely group differences and regression) recover the Average Treatment Effects defined in the previous section. Each simulation run represents a hypothetical SMART that begins by drawing 800 individual participants with a defined set of individual and experiment-specific parameters.

While the potential outcomes and causal estimands are purely design-based, we use models to provide structure to the simulations. In particular, many SMART designs track participants over time and the outcomes of interest represent a change from baseline. In these simulations, for simplicity, we assume that non-responders experience no change from where they would have been at baseline. However, this assumption does not affect our findings because the true causal effects are calculated within the simulation.

In our simulations, individuals are given an unobserved outcome level starting point, drawn from a standard normal distribution:

$$Y_{0i} \sim N(0,1)$$

The individual's potential outcomes for the intermediate tailoring variable $(I_i(A), I_i(B))$ are defined by their treatment assignment and their binary responder status associated with each

initial treatment $(R_{i,A}, R_{i,B})$. To carry through the principal strata from the previous section, these indicators determine the principal strata, as shown in Table 3.3 below.

Table 3.3

Principal Strata and Responder Indicators in Simulations

	$R_{i,B} = 0$	$R_{i,B} = 1$
$R_{i,A} = 0$	1. Never responders	2. Innovative responders
$R_{i,A} = 1$	3. Traditional responders	4. Always responders

We fit a parametric model to determine the potential outcomes for the intermediate tailoring variable as follows:

$$I_i(A) = Y_{0i} + \alpha_{i,A} R_{i,A}$$

$$I_i(B) = Y_{0i} + \alpha_{i,B} R_{i,B}$$

where (α_A, α_B) are parameters in the simulation and $\alpha_{i,B} \geq \alpha_{i,A}$. We then randomize individuals to their initial treatment status and "observe" their intermediate tailoring value. We then assign them Phase 2 potential outcomes depending on whether they are responders or non-responders. Again, for the purposes of the simulation, we use a parametric structure to determine the potential outcomes $Y_i(z, w)$, as shown in Table 3.4. As before, $(\alpha_{i,A}, \alpha_{i,B})$ are parameters relating to what happens during Phase 1 with $\alpha_{i,B} \geq \alpha_{i,A}$. We also parameterize what happens during Phase 2 with $(\delta_{i,1}, \delta_{i,2}, \delta_{i,A}, \delta_{i,B})$, where we assume $\delta_{i,1} \leq \delta_{i,2}$, $\alpha_{i,A} = \delta_{i,A}$, and $\alpha_{i,B} = \delta_{i,B}$ with no loss of generality.

Throughout all simulations, we vary several parameters within the model to reflect the standard SMART implementation in real-world experiments:

- $\bar{\alpha}_A$: the average additive effect for initial treatment A among responders
- $\bar{\alpha}_{i,B} - \bar{\alpha}_{i,A}$: the average difference between the additive effects of initial treatments A and B among Always Responders
- $\bar{\delta}_1$: the average additive effect for support 1 in Phase 2
- $\bar{\delta}_2 - \bar{\delta}_1$: the average difference between the additive effects of Supports 1 and 2 in Phase 2

Current simulations set the variance of $\bar{\alpha}_A$, $\bar{\alpha}_{i,B} - \bar{\alpha}_{i,A}$, $\bar{\delta}_1$ to zero, making these parameters constant across all individuals. The variance of $\bar{\delta}_2 - \bar{\delta}_1$ is sometimes larger than zero, as described more fully below.

Table 3.4

Parameterization of Simulated Potential Outcomes

	Non-responder to Treatment B	Responder to Treatment B
Non-responder to Treatment A	1. Never responders $Y_i(A,1) = Y_{0i} + \delta_{i,1}$ $Y_i(A,2) = Y_{0i} + \delta_{i,2}$ $Y_i(B,1) = Y_{0i} + \delta_{i,1}$ $Y_i(B,2) = Y_{0i} + \delta_{i,2}$	2. Innovative responders $Y_i(A,1) = Y_{0i} + \delta_{i,1}$ $Y_i(A,2) = Y_{0i} + \delta_{i,2}$ $Y_i(B,-) = Y_{0i} + \alpha_{i,B} + \delta_{i,B}$
Responder to Treatment A	3. Traditional responders $Y_i(A,-) = Y_{0i} + \alpha_{i,A} + \delta_{i,B}$ $Y_i(B,1) = Y_{0i} + \delta_{i,1}$ $Y_i(B,2) = Y_{0i} + \delta_{i,2}$	4. Always responders $Y_i(A,-) = Y_{0i} + \alpha_{i,A} + \delta_{i,A}$ $Y_i(B,-) = Y_{0i} + \alpha_{i,B} + \delta_{i,A}$

Finally, the other key individual parameter is their principal stratum. This study considers multiple different scenarios, corresponding to different data generating processes. Each scenario considers individual responsiveness (and thus principal strata) differently. Table 3.5 summarizes the different scenarios, which are described in more depth below.

Table 3.5

Simulation Scenarios

Scenario	Heterogeneity in Phase 2	Common Population Assumption
A-1	No	Met
A-2	No	Violated
B-1	Yes	Met
B-2	Yes	Violated

Scenarios A: Constant Phase 2 Treatment Effect

In Scenarios A-1 and A-2, we assume a constant Phase 2 treatment effect for both supports. However, we consider the case when the common population assumption is met and when it is violated. In Scenario A-1, the common population assumption means that individuals are either responsive to both Treatment A and B in Phase 1 or is responsive to neither Phase 1 treatment. We first draw a continuous, latent value of responsiveness for Treatment A and Treatment B from a multivariate normal, and then dichotomize those using quantiles of the normal distribution.

$$R_i^* \sim N(0,1)$$

$$R_i(A) = R_i(B) = (R_i^* > Q(1-p))$$

In Scenario A-2, the common population assumption is violated, and we draw two distinct measures of latent responsiveness in Phase 1. We first consider when these two values are independent, by setting their covariance to 0: responsiveness to Treatment A and B in Phase 1 are independent:

$$\begin{pmatrix} R_i^*(A) \\ R_i^*(B) \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

As before, the observed responsiveness variables are dichotomized using quantiles. In both cases, we simulate data across a range of values of p between 0 and 1.

$$R_i(A) = (R_i^*(A) > Q(1-p))$$

$$R_i(B) = (R_i^*(B) > Q(1-p))$$

Scenarios B: Correlation between Phase 1 Responsiveness and Phase 2 Effects

Scenario B relaxes the assumption of a constant Phase 2 treatment effect, to account for differential responses to additional supports. We operationalize this by inducing a correlation between a latent, continuous, Phase 1 responsiveness and the Phase 2 Effect. We then dichotomize the latent responsiveness into a binary indicator. In Scenario B-1, when the common population assumption holds, this is operationalized as follows:

$$\begin{pmatrix} R_i^* \\ \tau_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & r\sigma_\mu^2 \\ r\sigma_\mu^2 & \sigma_\mu^2 \end{pmatrix}\right)$$

$$R_i(A) = R_i(B) = (R_i^* > Q(1-p))$$

In this scenario, the responsiveness for Treatment A and B is dichotomized if latent responsiveness R_i^* is greater than the (1-p) quantile of the standard normal. The effect of Support 2 has a mean μ, a variance of σ_μ^2, and a correlation coefficient with latent responsiveness of r.

In Scenario B-2, we expand this model to a tri-variate normal with separate latent responsiveness values for Treatment A and Treatment B in Phase 1 that are uncorrelated. In both scenarios, we vary the values of μ, σ_μ^2, and r. In addition, we vary several experimental conditions as well. The experimental parameters include the effect of Treatments A and B among responders on the tailoring variable, and in general; and in the A- Scenarios, the constant effect of Phase 2 supports.

Using these individual and experiment parameters, individuals are categorized into one of the principal strata, and their potential outcomes are defined. Then, the simulation executes a SMART experiment by assigning individuals to Phase 1 treatments, assessing their responsiveness using a pre-set decision rule, and then assignment them to Phase 2 treatments when appropriate. Final observed outcomes are then created based on the individual's principal stratum assignment, their potential outcomes, and the realized treatment trajectory they received under the hypothetical design.

Analysis

The purpose of these simulations is to assess the average bias of the sample-driven estimators of the Phase 1 and Phase 2 main effects (from Equations 1 and 2). To do this, we calculate the bias in the first- and second-phase effect estimates by taking the difference between estimates from the observed data generated from our simulations and the corresponding true causal effect that we can calculate within each simulation. For simulation round *j*:

$$Bias_{j,1} = \widehat{ATE}_{j,1} - ATE_{j,1}$$

$$Bias_{j,2} = \widehat{ATE}_{j,2} - ATE_{j,2}$$

where $\widehat{ATE}_j$ is the regression coefficient on the Treatment B indicator. We take the mean bias ($\overline{Bias}$) across the different simulation runs and simulations parameters for each phase.

Additionally, we explore whether key parameters in the simulation predict the average bias across simulations.

Results

Bias in Phase 1 and Phase 2 Estimates

As shown in Figure 3.2, when the common population assumption holds, estimates of the first stage ATE appear to be unbiased with small amounts of estimation error. This means that when our estimates of the first-phase effects are only aggregating across the Always Responders and the Never Responders, we can recover the true ATE with some estimation noise. This is true regardless of whether the individual responses in Stage 2 are variable or constant. However, when the common population assumption is violated, and the sample includes Traditional Responders and Innovative Responders, we observe more bias in our estimates of the Phase 1 average treatment effect (ATE). This bias is small (ES< 0.005) when only 10% of the sample is identified as a responder after Phase 1 but is larger (ES >0.01) in other cases. Interestingly, the magnitude of the bias does not depend on the magnitude of the true ATEs, which in the simulations ranged from 0 to 0.5 standard deviations.

Figure 3.3 makes the same comparisons, but for estimating the Phase 2 average treatment effect (ATE). We see here that point estimates are upwardly biased in all cases, regardless of whether the common population assumption holds or not. Additionally, the bias is larger and much less predictable when there is a smaller number of responders in the sample – as the proportion of responders increases, the average bias approaches zero. As with the Phase 1 estimates, we seem to see equal amounts of bias regardless of whether the Stage 2 effects are constant or heterogeneous.

Figure 3.2

Bias in Estimating Phase 1 ATE

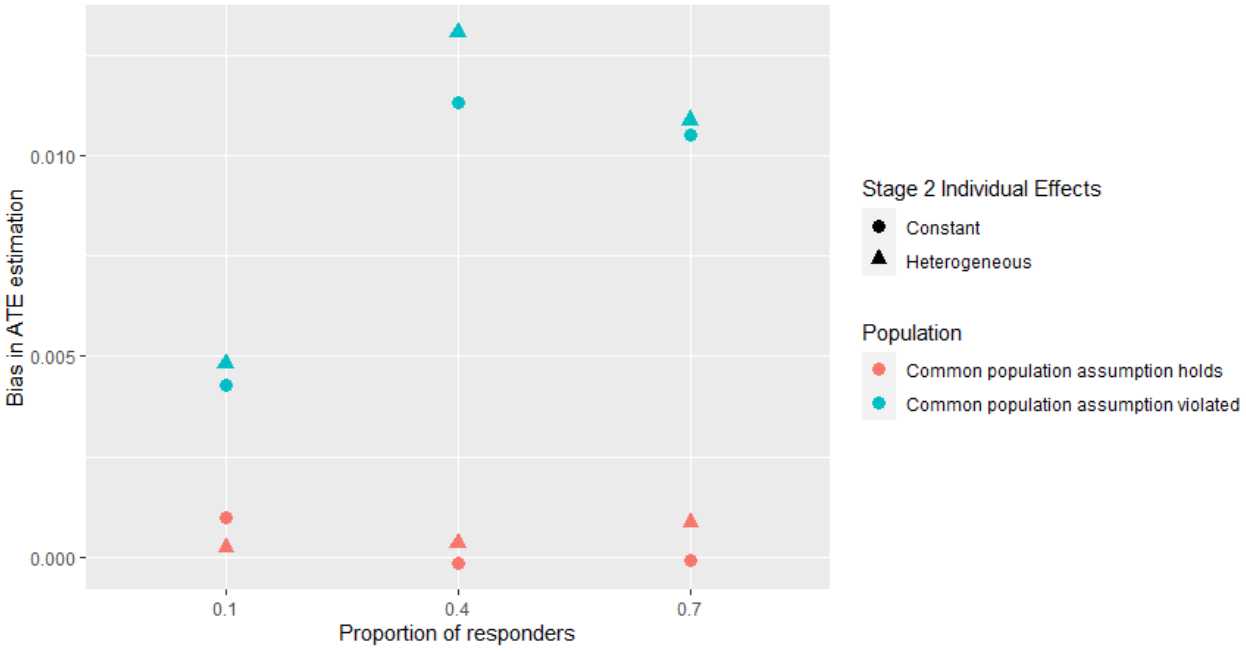
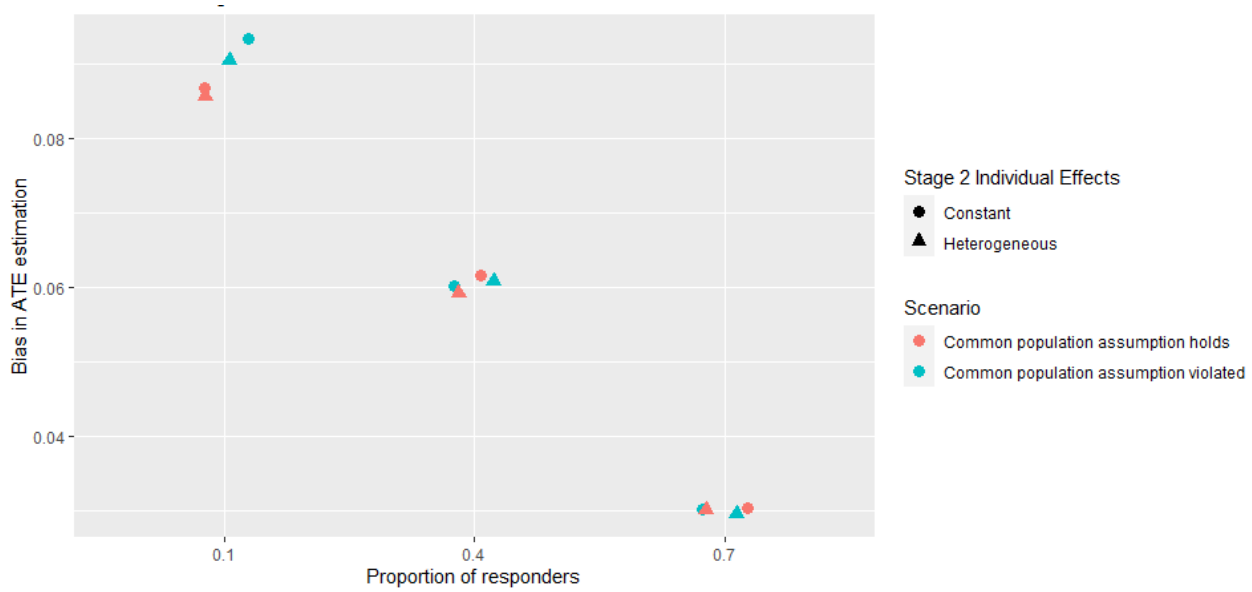


Figure 3.3

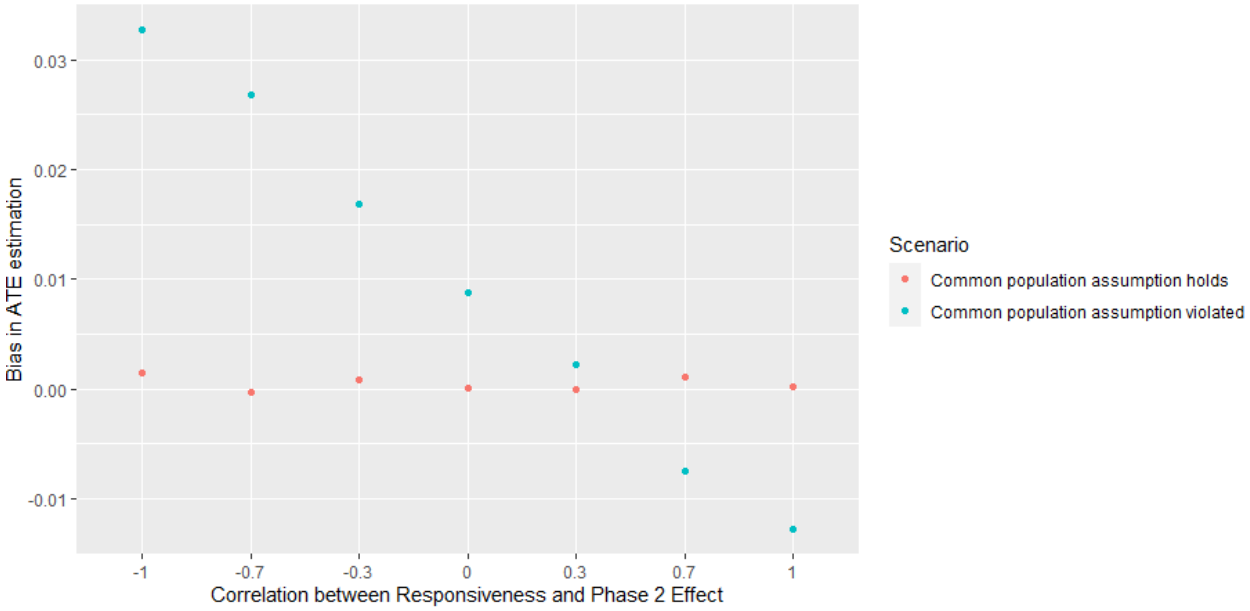
Bias in Estimating Phase 2 ATE



In the B scenarios, one of the conditions that we explicitly test is the extent to which correlation between individual responsiveness and the individual's effect in Phase 2 influences our inference. We display the results of this analysis below, in Figure 3.4 for the Phase 1 estimation and in Figure 3.5 for Phase 2 estimation. From these figures, we notice a few specific patterns. First, in Figure 3.4, we see that when the common population assumption holds, the average estimation bias holds for all correlation coefficients. However, when the common population assumption is violated, there is a strong negative correlation between the correlation coefficient and the average bias. A negative correlation is associated with a larger average bias, and a positive correlation is associated with a smaller and sometimes even negative average bias.

Figure 3.4

Bias in Estimating Phase 1 ATE with Heterogeneous Phase 2 Effects

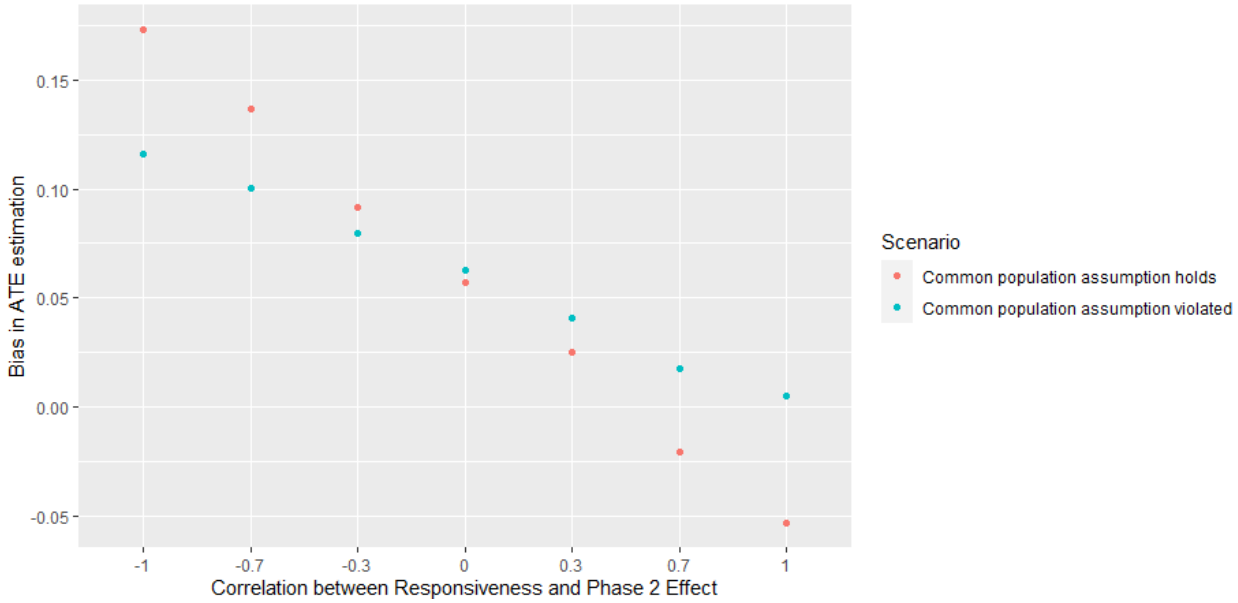


When we consider Phase 2 estimation in Figure 3.5, we again see a negative association between the correlation coefficient and the median bias, both when the common population assumption holds and when it is violated. In this case, however, the relationship is stronger when the assumption holds than when it is violated. Additionally, the magnitude of the average bias is

much larger in Phase 2 estimation than in Phase 1. Intuitively, this makes sense because the Phase 2 estimation is endogenous to what has already happened in Phase 1.

Figure 3.5

Bias in Estimating Phase 2 ATE with Heterogeneous Phase 2 Effects



Discussion

The purpose of this study is to define causal estimands of the Average Treatment Effect in the context of Sequential Multiple Assignment Randomized Trials (SMARTs) and explore the extent to which standard regression analysis can recover those estimands. To do this, we use principal stratification to identify four subpopulations: the Never Responders, who will not see improved outcomes regardless of what initial treatment they receive; the Always Responders, who will see improved outcomes regardless of what initial treatment they receive; the Traditional Responders, who will only see improved outcomes if they initially receive the traditional intervention; and Innovative Responders, who will only see improved outcomes if they initially receive the innovative intervention. We also show that the principal strata do not neatly align with the observed subgroups from our prototypical SMART design, displayed in Figure 3.1.

We use a common type of SMART design to show that the sets of available potential outcomes are different in each of these principal strata. By naming the potential outcomes, we can define the strata-specific individual treatment effects, the principal causal effects for each stratum, and the overall average treatment effect. We also identify the Common Population Assumption, which claims that individuals who do (or do not) respond to one initial treatment come from the same population as their counterparts who also respond (or do not) to the other initial treatment. In terms of our framework, this means that individuals belong to either the Never Responder or Always Responder principal strata. If there are Traditional Responders and Innovative Responders, the Common Population Assumption is violated, and knowing an individual's responsiveness in one initial treatment condition does not necessarily tell you about their responsiveness in the other initial treatment condition. Because the principal strata do not neatly align with the observed subgroups, this makes estimation based on observed subgroups risky.

Our simulation results show that estimating the Phase 1 and Phase 2 Average Treatment Effects using regression can result in non-zero bias. As shown in Figures 3.2 and 3.4, this bias in Phase 1 estimates is successfully eliminated when the Common Population Assumption holds. However, we believe it is unlikely to hold in many of the applications in which SMARTs are used. This is important, particularly since SMART designs are used to develop and refine interventions to be used in other contexts, where the population may nor may not be comparable to the population in the SMART. In cases where the populations are different, it might change the relative proportions of the different principal strata, which would result in different observed effects. However, understanding the principal causal effects would allow to easily predict how the average treatment effect might change under different conditions.

Additionally, the simulation reveals that when there is estimation bias, the magnitude of this bias is predicted by the correlation between an individual's responsiveness and their individual Phase 2 effect (Figure 3.4, 3.5). Strong negative correlations, or when being more responsive to the first phase intervention means that you will have a smaller Phase 2 effect, generate the largest positive bias. On the other hand, a strong positive correlation generates a negative bias (or an underestimate) of the Phase 2 ATE.

Limitations and Future Research

Beyond just investigating estimation bias in SMARTs, we plan to extend this work to look at coverage. Even if the estimates are biased, it is possible that the magnitude of the bias is sufficiently small to be captured within the 95% confidence intervals. Additionally, the current simulations make some simplifying assumptions. For example, while we allow latent responsiveness to each initial treatment to be continuous, we model it is as a binary indicator. Additional analyses could relax this assumption. A second assumption we make is that responsiveness to the assigned initial treatment condition is perfectly observed without measurement error. However, this is unlikely to be the case in applied settings, and future research should consider the implications of this measurement error, including "mislabeling" of responders and non-responders and implications for Phase 2 effectiveness.

This study also draws entirely on conceptual and simulation-based analysis. Future work would benefit from explicit analyses of how the different educational interventions and supports used in SMARTs to develop adaptive interventions behave. Specifically, even if we are skeptical that the Common Population Assumption will hold perfectly, we do not know how badly it will be violated in different educational contexts. If only a small proportion of the population is likely to fall in the Traditional Responder or Innovative Responder strata, that is better for causal

inference than a pair of initial interventions where almost the entire population falls into those two strata. Similarly, there is still much to learn about how initial responsiveness is related to responsiveness to the Phase 2 supports. Some of the 'optimization' analysis in SMARTs seeks to address this (for example, see Almirall et al., 2014), but current practices remain subject to the same concerns about biased inference that we address here. Finally, future research should explore both whether alternate estimators yield unbiased estimates of causal effects within SMARTs, for example Bayesian approaches for estimating principal causal effects (Mealli & Mattei, 2012) and whether those estimators are practical for broader use in education research.

In conclusion, the goal of this study is to apply a causal inference lens to the analysis of Sequential Multiple Assignment Randomized Trials. An important focus of this work was conceiving individual responsiveness to each initial treatment as a latent characteristic and using this characterization to define the four principal strata for which causal estimands are well defined. Applying the potential outcomes framework allows us to understand that standard sample-based estimators are making specific assumptions about the population. Our simulation study shows that when these assumptions are violated, our estimates of the Phase 1 and Phase 2 effects are more likely to be biased. Education scholars interested in using SMART designs should consider these findings when interpreting the results of their own SMART studies.

Conclusion

Treatment effect heterogeneity offers both challenges and opportunities for researchers and practitioners, and effective research design allows scholars to reveal and learn from this variation in impacts. Through three different studies, this dissertation highlights how a design that accommodates the presence of treatment effect heterogeneity can provide useful insights in the effectiveness of educational programs.

In the first paper, we show that slight differences in the design of a parental messaging intervention can yield different effects. While all message variations that we considered increased summer reading among families, the personalized messages were the most effective. Personalized messages also increased student test scores in the fall, relative to generic messages, but these effects depended on which view of reading was used to frame families' messages. Personalized messages were most effective when families received a mix of entertainment-framed messages and instrumentally framed messages and were least effective when families only received entertainment-framed messages. These findings contribute to a rapidly growing body of work showing that parental messages can be a useful tool for changing family reading behaviors and student reading outcomes.

The second paper finds that the overall null effects of a large, district-run universal pre-K program are masking substantial variation within the district. I show that students in the district who do not attending the district's pre-K program attend a variety of other early care options, including other subsidized programs like Head Start, private Georgia pre-K programs, and unsubsidized programs managed my for-profit and non-profit institutions. The areas where the district program shows no evidence of effectiveness are the same areas where the control group has the highest rates of pre-K attendance, particularly in unsubsidized programs. These control

group students also enter kindergarten more prepared than the control group students in other parts of the district where the district program is considered most effective. These findings show how universal-access programs interact with a broader set of early childhood educational opportunities and reveal the importance of considering the counterfactual when evaluating public pre-K programs.

The third paper highlights the importance of using traditional causal inference techniques when analyzing Sequential Multiple Assignment Randomized Trials. The principal strata reveal that that current analytic techniques have a latent assumption, which I call the Common Population Assumption, that individuals who respond well to one of the Phase 1 treatments within the SMART would have also responded to the other Phase 1 treatment. The assumption similarly states that those who do not respond to one of these initial treatments would also not have responded to the other. The simulation study shows that current estimates may not be recapturing an unbiased average treatment effect when this assumption is violated. As SMART designs become more popular in the space of education, this paper provides some guidance on the limitations of inferences researchers can draw from them.

Across all three studies, a common theme is that the presence of treatment effect heterogeneity requires consideration when designing educational research. In the third paper, the failure to consider that individuals respond differently to the two Phase 1 interventions is identified as an assumption that underlies common approaches to analyzing SMART designs. Violating this assumption can actually lead to estimation bias when attempting to study the effects of the two treatments. In the second paper, the failure to consider what pre-K experiences the control group receives will not cause biased estimates. But our inferences and interpretation of the program's effectiveness are shaped more by what is happening in the control group than

by what is happening in the district pre-K program. As a result, any implications and recommendations based on the analysis must also consider this counterfactual situation. And in the first paper, we show that a research design testing similar but distinct versions of a messaging campaign reveals more insight than could have been gained from a series of traditional randomized controlled trials. As education scholars continue the necessary research to support student learning, we must also continue to use research designs that consider the possibility of treatment effect heterogeneity.

Appendices: Additional Materials

Table A.1

Text Message Topics, Organized by Theme

Summer Reading Resources	**Reminders to Engage in Reading Activities**	**Monitoring progress**
· App feature: availability of books · App feature: personalized activities/goal reminder · App feature: "catching words" · Resource: chromebooks available at library · App feature: new content · App feature: 2 sets of activities	· Tip: talking about books · App is great · Tip: planning time to read · Social pressure: number of users · Tip: talk about favorite book	· Kickoff & goal setting · Monitoring: check-in on goals/progress · Monitoring: check-in on goals/progress · Closeout & survey preview

Table A.2

Comparing Message Variations Across Conditions for Two Example Messages

	Generic	Personalized information
Message 1		
Instrumental view	The MORE@Home app contains personalized activities for each book that will help [StudentFirstName] **develop reading skills**.	The MORE@Home app contains personalized activities for each book. [StudentFirstName] can **practice different reading skills** for both
Entertainment view	The MORE@Home app contains personalized activities for each book. We think [StudentFirstName] **will have fun doing them all!**	The MORE@Home app contains personalized activities for each book. [StudentFirstName] can **have fun exploring them** for both **[Book1Title] and [Book2Title]**.
Message 2		
No goals	We are already 6 weeks into summer vacation! Hopefully you and [StudentFirstName] are making progress on your summer reading list!	We are 6 weeks into summer vacation and [StudentFirstName] **has used [NumBooksAccessed] books on the MORE@Home app**. Keep reading to get through all of them!
Goals	We're already 6 weeks into summer vacation! At this point, you should be about **1/2 of the way to your summer reading goal**.	We're already 6 weeks into summer vacation! At this point, you should be about 1/2 of your way to your **summer reading goal of [BookGoal] books**.

Figure B. 1

Flow Chart Showing Construction of Analytic Samples, Paper 2

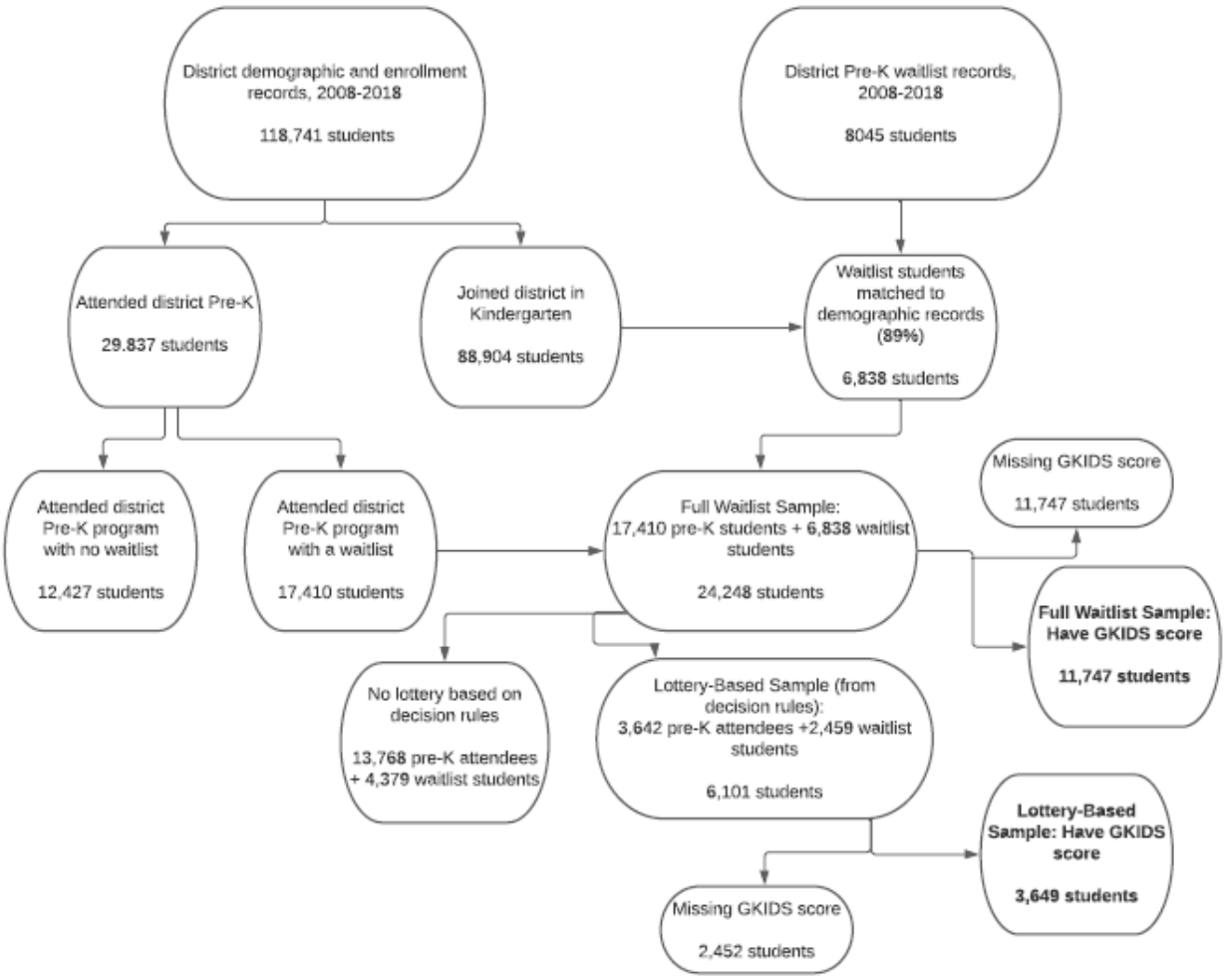


Table B.1

Demographic Characteristics for Students with Known and Unknown Pre-K Experiences

	Full Waitlist Sample		
Characteristic	**Known Pre-K Experience Group Mean**	**Unknown Pre-K Experience Group Mean**	**Difference**
Male (%)	50.2	51.2	-1.0
Race			
African American (%)	46.6	47.5	-0.9 *
Hispanic (%)	15.9	16.2	-0.3
White (%)	25.0	23.6	1.4 **
Asian (%)	9.6	9.7	-0.1
Other (%)	3.0	3.0	0.0
English Language Learner (%)	13.7	13.8	-0.1 **
FRL-eligible (%)	54.1	53.4	0.7
Have Individualized Education Plan (IEP) (%)	8.8	8.1	0.7
Number of Students	35,342	37,868	

Notes: *p <0.05, **p<0.01, ***p<0.001

References

Almirall, D., Kasari, C., McCaffrey, D. F., & Nahum-Shani, I. (2018). Developing optimized adaptive interventions in education. *Journal of Research on Educational Effectiveness*, *11*(1), 27–34. https://doi.org/10.1080/19345747.2017.1407136

Almirall, D., Nahum-Shani, I., Sherwood, N. E., & Murphy, S. A. (2014). Introduction to SMART designs for the development of adaptive interventions: with application to weight loss research. *Translational Behavioral Medicine*, *4*(3), 260–274. https://doi.org/10.1007/s13142-014-0265-0

Athey, S., & Imbens, G. W. (2017). Chapter 3—The econometrics of randomized experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 1, pp. 73–140). North-Holland. https://doi.org/10.1016/bs.hefe.2016.10.003

Baker, L., Mackler, K., Sonnenschein, S., & Serpell, R. (2001). Parents' interactions with their first-grade children during storybook reading and relations with subsequent home reading activity and reading achievement. *Journal of School Psychology*, *39*(5), 415–438. https://doi.org/10.1016/S0022-4405(01)00082-6.

Baker, L., & Scher, D. (2002). Beginning readers' motivation for reading in relation to parental beliefs and home reading experiences. *Reading Psychology*, *23*(4), 239–269. https://doi.org/10.1080/713775283.

Baker, L., Scher, D., & Mackler, K. (1997). Home and family influences on motivations for reading. *Educational Psychologist*, *32*(2), 69–82. https://doi.org/10.1207/s15326985ep3202_2.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2016). Evaluation of response to intervention practices for elementary school reading (NCEE 2016-4000). Washington, D.C.: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://files.eric.ed.gov/fulltext/ED560820.pdf

Barnett, W.S., Jung, K., Friedman-Krauss, A., Frede, E. C., Nores, M., Hustedt, J. T., Howes, C., & Daniel-Echols, M. (2018). State prekindergarten effects on early learning at kindergarten entry: An analysis of eight state programs. *AERA Open, 4*(2), 1-16. https://doi.org/10.1177/2332858418766291

Bell, S. H., Olsen, R. B., Orr, L. L., & Stuart, E. A. (2016). Estimates of external validity bias when impact evaluations select sites nonrandomly. *Educational Evaluation and Policy Analysis*, *38*(2), 318–335. https://doi.org/10.3102/0162373715617549

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2015). Turning a shove into a nudge? A "labeled cash transfer" for education. *American Economic Journal: Economic Policy*, *7*(3), 86–125. https://doi.org/10.1257/pol.20130225

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bird, K. A., Castleman, B. L., Denning, J. T., Goodman, J., Lamberton, C., & Rosinger, K. O. (2021). Nudging at scale: Experimental evidence from FAFSA completion campaigns. *Journal of Economic Behavior & Organization, 183*, 105-128. https://doi.org/10.1016/j.jebo.2020.12.022

Bloom, H. S., & Unterman, R. (2014). Can small high schools of choice improve educational prospects for disadvantaged students. *Journal of Policy Analysis and Management*, *33*(2), 290–319. https://doi.org/10.1002/pam.21748

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. MDRC. https://doi.org/10.2139/ssrn.2594430

Cabell, S. Q., Zucker, T. A., DeCoster, J., Copp, S. B., & Landry, S. (2019). Impact of a parent text messaging program on pre-kindergarteners' literacy development. *AERA Open*, *5*(1), 1-16. https://doi.org/10.1177/2332858419833339

Cascio, E. U. (2017). Does universal preschool hit the target? Program access and preschool impacts (No. w23215). National Bureau of Economic Research. https://doi.org/10.3386/w23215

Castleman, B. L. & Page, L. C. (2015). Summer nudging: Can personalized text messages and peer mentor outreach increase college going among low-income high school graduates?" *Journal of Economic Behavior & Organization*, *115*, 144-160. https://doi.org/10.1016/j.jebo.2014.12.008.

Castleman, B.L. & Page L.C. (2016). Freshman year financial aid nudges: An experiment to increase FAFSA renewal and college persistence." *The Journal of Human Resources*, *51*(2), 389- 415. https://doi.org/ 10.3368/jhr.51.2.0614-6458R.

Collins, L. M., Dziak, J. J., Kugler, K. C., & Trail, J. B. (2014). Factorial experiments: Efficient tools for evaluation of intervention components. *American Journal of Preventive Medicine*, *47*(4), 498–504. https://doi.org/10.1016/j.amepre.2014.06.021

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent ehealth interventions. *American Journal of Preventive Medicine, 32*(5, Supplement), S112–S118. https://doi.org/10.1016/j.amepre.2007.01.022

Condliffe, B., & Balu, R. (2019). Missing from the start: Engagement in New York City's kindergarten application. MDRC. https://www.mdrc.org/publication/missing-start

Condliffe, B. F., Boyd, M. L., & DeLuca, S. (2015). Stuck in school: How social context shapes school choice for inner-city students. *Teachers College Record, 117*(3).

Connor, C. M. (2017). Using technology and assessment to personalize instruction: Preventing reading problems. *Prevention Science*, *20*(1), 89-99. https://doi.org/10.1007/s11121-017-0842-9

Connor, C. M., & Morrison, F. J. (2016). Individualizing student instruction in reading: implications for policy and practice. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 54–61. https://doi.org/10.1177/2372732215624931

Cortes, K.E, Fricke, H., Loeb, S., Song, D., & York, B. (2019). When behavioral barriers are too high or low – how timing matters for parenting interventions (No. w25964; p. w25964). National Bureau of Economic Research. https://doi.org/10.3386/w25964

Cortes, K. E., Fricke, H., Loeb, S., Song, D. S., & York, B. N. (2021). Too little or too much? Actionable advice in an early-childhood text messaging experiment. *Education Finance and Policy, 16*(2), 1–44. https://doi.org/10.1162/edfp_a_00304

DellaVigna, S., & Linos, E. (2020). RCTs to scale: Comprehensive evidence from two nudge units (No. w27594). National Bureau of Economic Research. https://doi.org/10.3386/w27594

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics, 1*(3), 111-134. http://www.aeaweb.org/articles.php?doi=10.1257/app.1.3.111

Denton, C. A., Kethley, C., Nimon, K., Kurz, T. B., Mathes, P. G., Minyi Shih, & Swanson, E. A. (2010). Effectiveness of a supplemental early reading intervention scaled up in multiple schools. *Exceptional Children*, *76*(4), 394–416. https://doi.org/10.1177/001440291007600402

Doss, C., Fahle, E. M., Loeb, S., & York, B. N. (2019). More than just a nudge: Supporting kindergarten parents with differentiated and personalized text messages. *Journal of Human Resources*, *54*(3), 567–603. https://doi.org/10.3368/jhr.54.3.0317-8637R

Doss, C., Fricke, H., and Loeb, S. (2020). Math is for girls: The unequal effects of text messaging to help parents support early math development. (EdWorkingPaper: 20-310). Annenberg Institute at Brown University. https://doi.org/10.26300/39zc-j672.

Duncan, G. J., & Vandell, D. L. (2012). A conceptual approach to understanding treatment heterogeneity in human capital interventions. Society for Research on Educational Effectiveness. Retrieved from https://eric.ed.gov/?id=ED530411

Early, D. M., Li, W., Maxwell, K. L., & Ponder, B. D. (2019). Participation in Georgia's pre-k as a predictor of third-grade standardized test scores. *AERA Open, 5*(2), 1-16. https://doi.org/10.1177/2332858419848687

Feller, A., Grindal, T., Miratrix, L., Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics, 10*(3): 1245–1285. https://doi.org/10.1214/16-aoas910

Fellers, L. A. (2017). Developing an approach to determine generalizability: A review of efficacy and effectiveness trials funded by the Institute of Education Sciences [Doctoral dissertation, Columbia University]. https://doi.org/10.7916/D86D5ZN1

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*(1), 21–29. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4137767/

Georgia Department of Early Care and Learning. (2018). Georgia's pre-k program public bulletin: 2016-2017 school year. Retrieved from www.decal.ga.gov.

Georgia Department of Education. (n.d.). GKIDS Readiness Check. Retrieved from https://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Pages/Readiness.aspx.

Gormley, W. T., & Phillips, D. (2005). The effects of universal pre-K in Oklahoma: Research highlights and policy implications. *Policy Studies Journal*, *33*(1), 65–82. https://doi.org/10.1111/j.1541-0072.2005.00092.x

Hardy, M. (1993). Regression with dummy variables. SAGE Publications, Inc. https://doi.org/10.4135/9781412985628

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902. https://doi.org/10.1126/science.1128898

Heppen, J., Kurki, A., & Brown, S. (2020). *Can Texting Parents Improve Attendance in Elementary School? A Test of an Adaptive Messaging Strategy (NCEE 2020–006a)* (p. 134). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. http://ies.ed.gov/ncee.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hollands, F. M., Kieffer, M. J., Shand, R., Pan, Y., Cheng, H., & Levin, H. M. (2016). Cost-effectiveness analysis of early reading programs: A demonstration with recommendations for future research. *Journal of Research on Educational Effectiveness, 9*(1), 30–53. https://doi.org/10.1080/19345747.2015.1055639

Hurwitz, L. B., Lauricella, A. R., Hanson, A., Raden, A., & Wartella, E. (2015). Supporting Head Start parents: Impact of a text message intervention on parent–child activity engagement. *Early Child Development and Care, 185*(9), 1373–1389. https://doi.org/10.1080/03004430.2014.996217

Ideas42. (2015a). Increasing FAFSA applications: Making college more affordable. New York. Retrieved from http://www.ideas42.org/wp-content/uploads/2015/12/FAFSA-Brief.pdf

Ideas42. (2015b). Choosing courses to stay eligible to financial aid. New York. Retrieved from http://www.ideas42.org/wp-content/uploads/2015/12/Valencia-Brief.pdf

Imai, K., King, G., & Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *171*(2), 481–502. https://doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G.W. & Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction,* Chapters 14-16. New York: Cambridge University Press.

Iowa Department of Education. (2020). Statewide voluntary preschool program for four-year-old children fact sheet. https://educateiowa.gov/sites/files/ed/documents/SWVPPFactSheet-Fall2020.pdf

Jacob, R. T., Armstrong, C., Bowden, A. B., & Pan, Y. (2016). Leveraging volunteers: An experimental evaluation of a tutoring program for struggling readers. *Journal of Research on Educational Effectiveness*, *9*(S1), 67–92.

Jones, J. S., Conradi, K., & Amendum, S. J. (2016). Matching interventions to reading needs: A case for differentiation. *The Reading Teacher*, 70(3), 307–316. https://doi.org/10.1002/trtr.1513

Joseph, L. M. (2008). Best practices on interventions for students with reading problems. *National Association of School Psychologists Communique*, *37*(4), 12-13.

Kidwell, K. M., & Hyde, L. W. (2016). Adaptive interventions and SMART designs: Application to child behavior research in a community setting. *The American Journal of Evaluation*, *37*(3), 344–363. https://doi.org/10.1177/1098214015617013

Kim, J. S., Asher, C. A., Burkhauser, M., Mesite, L., & Leyva, D. (2019). Using a sequential multiple assignment randomized trial (SMART) to develop an adaptive K–2 literacy intervention with personalized print texts and app-based digital activities. *AERA Open* *5*(3), 1-18. https://doi.org/10.1177/2332858419872701

Kim, J. S., Guryan, J., White, T. G., Quinn, D. M., Capotosto, L., & Kingston, H. C. (2016). Delayed effects of a low-cost and large-scale summer reading intervention on elementary school children's reading comprehension. *Journal of Research on Educational Effectiveness*, *9*(S1), 1–22. https://doi.org/10.1080/19345747.2016.1164780

Kline, P., & Walters, C.R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics, 131*(4), 1795–1848. https://doi.org/10.1093/qje/qjw027

Kraft, M. A., & Monti-Nussbaum, M. (2017). Can schools enable parents to prevent summer learning loss? A text-messaging field experiment to promote literacy skills. *The ANNALS of the American Academy of Political and Social Science*, *674*(1), 85–112. https://doi.org/10.1177/0002716217732009

Kugler, K. C., Trail, J. B., Dziak, J. J., & Collins, L. M. (2012). Effect coding versus dummy coding in analysis of data from factorial experiments. The Methodology Center,

Pennsylvania State University. Retrieved from https://www.methodology.psu.edu/

Ludwig J. & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*(1), 159-208.

Lynch, J., Anderson, J., Anderson, A., & Shapiro, J. (2006). Parents' beliefs about young children's literacy development and parents' literacy behaviors. *Reading Psychology*, 27(1), 1–20. https://doi.org/10.1080/02702710500468708

Marinak, B. A., Malloy, J. B., Gambrell, L. B., & Mazzoni, S. A. (2015). Me and My Reading Profile: A tool for assessing early reading motivation. *The Reading Teacher*, 69(1), 51–62. https://doi.org/10.1002/trtr.1362

Mayer, S. E., Kalil, A., Oreopoulos, P., & Gallegos, S. (2018). Using behavioral insights to increase parental engagement: The Parents and Children Together intervention. *Journal of Human Resources*, *54*(4), 900-925. https://doi.org/10.3368/jhr.54.4.0617.8835R

McCormick, M., Mattera, S., & Hsueh, J. (2019). Preschool to third grade alignment: What do we know and what are we learning?. New York: MDRC. https://www.mdrc.org/publication/preschool-third-grade-alignment

McCoy, D. C., Morris, P. A., Connors, M. C., Gomez, C. J., & Yoshikawa, H. (2016). Differential effectiveness of Head Start in urban and rural communities. *Journal of Applied Developmental Psychology*, *43*, 29–42. https://doi.org/10.1016/j.appdev.2015.12.007

Mealli, F., & Mattei, A. (2012). A Refreshing Account of Principal Stratification. *The International Journal of Biostatistics*, *8*(1). https://doi.org/10.1515/1557-4679.1380

Northwest Evaluation Association (NWEA). (2011). Measure of Academic Progress.

Muralidharan, K., Romero, M., & Wüthrich, K. (2019). Factorial designs, model selection, and (incorrect) inference in randomized experiments (No. w26562). National Bureau of Economic Research. https://doi.org/10.3386/w26562.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J., & Murphy, S. A. (2012). Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological Methods*, *17*(4), 457–477. https://doi.org/10.1037/a0029372

Nation's Report Card. (2015). *Nation's Report Card*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. http://www.nationsreportcard.gov/reading_math_2015/#reading?grade=4

Nickerson, D. W., & Rogers, T. (2010). Do you have a voting plan?: Implementation intentions, voter turnout, and organic plan making. *Psychological Science*, *21*(2), 194–199. https://doi.org/10.1177/0956797609359326

Oettingen, G., & Reininger, K. M. (2016). The power of prospection: Mental contrasting and behavior change. *Social and Personality Psychology Compass, 10*(11), 591–604. https://doi.org/10.1111/spc3.12271

Oreopoulos, P., Petronijevic, U., Logel, C., & Beattie, G. (2020). Improving non-academic student outcomes using online and text-message coaching. *Journal of Economic Behavior & Organization, 171*, 342–360. https://doi.org/10.1016/j.jebo.2020.01.009

Owens, A., Reardon, S. F., & Jencks, C. (2016). Income segregation between schools and School districts. *American Educational Research Journal, 53*(4), 1159–1197. https://doi.org/10.3102/0002831216652722

Page, L. C., Castleman, B. L., & Meyer, K. (2020). Customized nudging to improve FAFSA completion and income verification. *Educational Evaluation and Policy Analysis*, *42*(1), 3-21. https://doi.org/10.3102/0162373719876916.

Parker, E., Diffey, L., & Atchison, B. (2018). How states fund pre-K: A primer for policymakers. Education Commission of the States. https://www.ecs.org/wp-content/uploads/How-States-Fund-Pre-K_A-Primer-for-Policymakers.pdf

Pearman, F. A. (2020). The moderating effect of neighborhood poverty on preschool effectiveness: evidence from the Tennessee voluntary prekindergarten experiment. *American Educational Research Journal, 57*(3), 1323–1357. https://doi.org/10.3102/0002831219872977

Peisner-Feinberg, E. S., & Schaaf, J. M. (2011). Effects of the North Carolina More at Four pre-kindergarten program on children's school readiness skills. Key Findings. Chapel Hill: The University of North Carolina, FPG Child Development Institute. https://files.eric.ed.gov/fulltext/ED598158.pdf

Peisner-Feinberg, E. S., Schaaf, J. M., LaForett, D. R., Hildebrandt, L.M., & Sideris, J. (2014). Effects of Georgia's pre-K program on children's school readiness skills: Findings from the 2012–2013 evaluation study. Chapel Hill: The University of North Carolina, FPG Child Development Institute. https://fpg.unc.edu/sites/fpg.unc.edu/files/resources/reports-and-policy-briefs/GAPreKEval_RDDReport%203-4-2014.pdf

Pelham, W. E., Fabiano, G. A., Waxmonsky, J. G., Greiner, A. R., Gnagy, E. M., Pelham, W. E., Coxe, S., Verley, J., Bhatia, I., Hart, K., Karch, K., Konijnendijk, E., Tresco, K., Nahum-Shani, I., & Murphy, S. A. (2016). Treatment sequencing for childhood ADHD: A multiple-randomization study of adaptive medication and behavioral interventions. *Journal of Clinical Child & Adolescent Psychology*, *45*(4), 396–415. https://doi.org/10.1080/15374416.2015.1105138

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M.R., Duncan, G. J., Dynarski, M., Magnuson, K.A., Weiland, C. (2017). Puzzling it out: the current state of scientific knowledge on pre-kindergarten effects. Washington, D.C.: Brookings Institute. https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

Robinson, C. D., Lee, M. G., Dearing, E., & Rogers, T. (2018). Reducing student absenteeism in the early grades by targeting parental beliefs. *American Educational Research Journal*, *55*(6), 1163–1192. https://doi.org/10.3102/0002831218772274

Rogers, T., & Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour, 2*(5), 335–342. https://doi.org/10.1038/s41562-018-0328-1

Rogers, T., Duncan, T., Wolford, T., Ternovski, J., Subramanyam, S., & Reitano, A. (2017). A randomized experiment using absenteeism information to "nudge" attendance (REL 2017-252). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. http://ies.ed.gov/ncee/edlabs

Rogers, T., Milkman, K. L., John, L. K., & Norton, M. I. (2015). Beyond good intentions: Prompting people to make plans improves follow-through on important tasks. *Behavioral Science & Policy*, *1*(2), 33–41. https://doi.org/10.1353/bsp.2015.0011

Rojas, N. M., Morris, P., & Balaraman, A. (2020). Finding rigor within a large-scale expansion of preschool to test impacts of a professional development program. *AERA Open, 6*(4), 1-21. https://doi.org/10.1177/2332858420975399

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1080/19345747.2016.1138560

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities*, *48*(4), 369–390. https://doi.org/10.1177/0022219413504995

Schanzenbach, D. Whitmore. (2006). What have researchers learned from project STAR? *Brookings Papers on Education Policy 2006*, 205–228. https://doi.org/10.1353/pep.2007.0007

Schochet, P. Z., Puma, M., & Deke, J. (2014). Understanding variation in treatment effects in education impact evaluations: An overview of quantitative methods (NCEE 2014-4017). Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Analytic Technical Assistance and Development. https://ies.ed.gov/ncee/pubs/20144017/

Shapiro, A., Martin, E., Weiland, C., & Unterman, R. (2019). If you offer it, will they come? Patterns of application and enrollment behavior in a universal prekindergarten context. *AERA Open*, 5(2), 1-22. https://doi.org/10.1177/2332858419848442

Smythe-Leistico, K., & Page, L. C. (2018). Connect-text: Leveraging text-message communication to mitigate chronic absenteeism and improve parental engagement in the earliest years of schooling. *Journal of Education for Students Placed at Risk (JESPAR)*, *23*(1–2), 139–152. https://doi.org/10.1080/10824669.2018.1434658

Somers, M.-A., Collins, L., & Maier, M. (2014). Design options for an evaluation of head start coaching: Review of methods for evaluating components of social interventions (OPRE Report #2014-81; Produced by American Institutes for Research for Head Start Professional Development: Developing Evidence for Best Practices in Coaching.). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation. https://www.mdrc.org.

Sonnenschein, S., Baker, L., Serpell, R., Scher, D., Fernandez-Fein, S., & Munsterman, K. A. (1996). Strands of emergent literacy and their antecedents in the home: Urban preschoolers early literacy development. National Reading Research Center. https://files.eric.ed.gov/fulltext/ED392019.pdf

Sonnenschein, S., Baker, L., Serpell, R., Scher, D., Truitt, V. G., & Munsterman, K. (1997). Parental beliefs about ways to help children learn to read: The impact of an entertainment or a skills perspective. *Early Child Development and Care*, 127(1), 111–118. https://doi.org/10.1080/0300443971270109

Sonnenschein, S., Baker, L., Serpell, R., & Schmidt, D. (2000). Reading is a source of entertainment: The importance of the home perspective for children's literacy development. In K. A. Roskos & J. F. Christie (Eds.), *Play and literacy in early childhood: Research from multiple perspectives* (pp. 107–124). Lawrence Erlbaum Associates Publishers.

U.S. Department of Health and Human Services, Administration for Children and Families (2010). Head Start Impact Study. Final Report. Washington, DC. https://files.eric.ed.gov/fulltext/ED507845.pdf

Unterman, R., & Weiland, C. (2020). Higher-Quality Elementary Schools Sustain the Prekindergarten Boost: Evidence from an Exploration of Variation in the Boston Prekindergarten Program's Impacts. In EdWorkingPapers.com. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai20-321

van Huizen, T., & Plantenga, J. (2018). Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments. *Economics of Education Review, 66*, 206–222. https://doi.org/10.1016/j.econedurev.2018.08.001

Weiland, C., Unterman, R., Shapiro, A., Staszak, S., Rochester, S., & Martin, E. (2020). The effects of enrolling in oversubscribed prekindergarten programs through third grade. *Child Development*, *91*(5), 1401–1422. https://doi.org/10.1111/cdev.13308

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84(*6), 2112- 2130.

Weiss, M. J., Bloom, H., S., & Brock, T. (2014). A conceptual framework for studying the variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876. https://doi.org/10.1080/19345747.2017.1300719

West Virginia Department of Education, Office of Early Learning. (2016). Early Learning Annual Report. https://files.eric.ed.gov/fulltext/ED594492.pdf

Willingham, D. T. (2021, March 2). Making education research relevant. *Education Next*. https://www.educationnext.org/making-education-research-relevant-how-researchers-can-give-teachers-more-choices/

York, B. N., Loeb, S., & Doss, C. (2019). One step at a time: The effects of an early literacy text messaging program for parents of preschoolers. *Journal of Human Resources, 54*(3), 537-566. https://doi.org/10.3368/jhr.54.3.0517-8756R