Improving Teacher Supports through Unconventional Experiments

---

A Dissertation

Presented to

The Faculty of the Curry School of Education

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Katherine L. Miller-Bains, B.A., M.A.

December 18, 2018

Abstract

Vivian Wong, Advisor

Teachers are vital to students' learning and success. Yet, we have little evidence of how to best support teachers and their practice, particularly as professional demands shift to accommodate more technology, greater diversity, and higher standards for learning. The ability to make causal links between professional supports and teacher and student outcomes is further constrained by the limited methods typically used by researchers. Because traditional experiments are often long, expensive, and politically infeasible, they are reserved for later stages of evaluation. This dissertation presents three papers, each applying an underutilized method for researching and developing effective supports for pre- and in-service teachers, with a particular focus on supports to help educators to understand and make use of student data. The first study uses an alternative-treatment experimental design in order to investigate the impacts of a brief, low-cost data literacy workshop within a teacher preparation program. Paper 2 evaluates a scalable intervention to help kindergarten teachers use entry assessment data to inform instruction using an embedded factorial design. Finally, Paper 3 describes a special repeated measures design – the experimental switching replication – as well as how it was used to evaluate skills-based coaching in a simulated learning environment. Together, these papers illustrate different approaches to program evaluation, highlighting ways in which underutilized research methods can address the design and development of teacher supports.

Department of Educational Leadership, Foundations, and Policy
Research, Statistics, and Evaluation
Curry School of Education
University of Virginia
Charlottesville, VA

APPROVAL OF THE DISSERTATION

The dissertation ("Improving Teacher Supports through Unconventional Experiments") has been approved by the Graduate Faculty of the Curry School of Education in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Name of Chair (Dr. Vivian C. Wong)

_____
Committee Member (Dr. Julie J. Cohen)

_____
Committee Member (Dr. Bridget Hamre)

_____
Committee Member (Dr. Amanda Williford)

Date: _____

Dedication

To Adam, for being the most thoughtful, caring partner, and keeping me fed and grounded throughout this process.

To Gay and Mike Miller, who have always believed that I was capable of more than I realized.

To the friends, family, and educators in my life, whom I have been fortunate to have as role models, inspirations, and instigators.

Acknowledgements

Thank you so much to Vivian Wong for being my friend and mentor throughout my time in Charlottesville. Thank you for always taking the time to provide guidance and support – both professionally and personally.

Thank you to Amanda Williford for giving me the chance to join her team five years ago and continuing to provide me with advice, mentoring, and opportunities.

Thank you to Julie Cohen for always caring and letting me help to make your sim-dreams into a (simulated) reality.

Thank you to Bridget Hamre for agreeing to be on my dissertation committee, even in the midst of a big transition, and always providing thoughtful and important feedback.

Thank you to Jason Downer, Genna Matthew, Elise Rubinstein, Jaclyn Russo, Jessica Vick Whittaker, and Wanda Weaver the entire Virginia Kindergarten Readiness team, who have provided so much support to me and, more importantly, hundreds of teachers across the state.

Thank you to Anna Markowitz for the numerous coffee runs, free counseling, and thoughtful conversations. We will always have Gomey and Claremont.

Thank you to Anandita, Emily, Kelly, Kylie, and Rebekah for being thought partners and partners in crime.

Thank you to Beth Tipton for encouraging me to pursue this path and for making sure Vivian and I found each other.

TABLE OF CONTENTS

LIST OF TABLES

TABLE Page

LIST OF FIGURES AND APPENDICES

FIGURE Page

Study 1

Study 2

Improving Teacher Supports through Unconventional Experiments

Katherine L. Miller-Bains

University of Virginia

Conceptual Link: Improving Teacher Supports through Unconventional Experiments

Over the past two decades, federal agencies such as the Institute for Education Sciences and the U.S. Department of Education have prioritized the funding of experimental trials in education research. As a result, an increasing number of educational studies rely on experiments to provide impact estimates for interventions and programs (Stockard & Wood, 2016). The increased prevalence of such experimental research has improved our understanding of what works and what does not in education. Yet, we still have limited information about how to best support educators as they progress through preparation programs and into the classroom (Jacob & McGovern, 2016). Despite the limited evidence on returns from teacher professional supports to student outcomes, departments of education across the country continue to make large investments in programs intended to support teacher learning.

The current literature base on teacher professional learning is mixed. Causal research has connected changes in teachers' practice as result of effective professional learning to improvements in student learning (Kraft, Blazer, & Hogan, 2018; Scher & O'Reilly, 2009). However, this connection appears tenuous, as large-scale evaluations of teacher professional development find that – on average – the formal supports and training teachers receive are not associated with gains in student outcomes (Garet, et al., 2016; Harris & Sass, 2011). Collectively, this evidence suggests that while professional learning experiences are able to improve educators' knowledge, skills, and practices in ways that impact students, teachers are not consistently exposed to high-quality training on a large scale. The variability in professional supports for teachers can be partly

attributed to the high cost associated with many programs that have been identified as effective – such as the use of instructional coaches (Kraft, Blazer, & Hogan, 2018) – as well as the difficulty in effectively scaling up interventions and supports (Coburn, 2003). As such, the education community needs rigorous research to support the evaluation and development of scalable teacher development.

Although experiments can fill the need for additional causal evidence on the effectiveness of teacher supports, traditional randomized-control trials (RCTs) are difficult to execute in educational settings. First, they require withholding treatment from a subset of the sample, making them politically infeasible in many settings. Additionally, they are practically difficult to carry out, as educational studies are frequently lengthy, draw from limited samples, and evaluate treatments with small effects. This is particularly true when teachers are the unit of analysis, increasing the concern for spillover effects and further constraining the statistical power of a study. Finally, RCTs answer limited questions about the efficacy of interventions, as they evaluate multi-faceted interventions against a control group within constrained samples and settings. As such, experiments are typically reserved for late stages of research and development, when program developers have less latitude to address implementation issues. Furthermore, these late-stage randomized controlled trials provide little insight into which intervention features are necessary in order to achieve the desired effects.

This three-paper dissertation seeks to both improve the evidence on effective teacher supports and expand the experimental approaches employed by educational researchers. I highlight alternative experimental designs such as the factorial experiment and switching replication, which can enable researchers to answer a wider array of

research questions while also combatting issues common in educational studies such as small sample sizes and political feasibility. Furthermore, expanding the types of experiments employed in education can allow researchers to better identify causal mechanisms – allowing the educational community to build more efficient and effective interventions. Each chapter of this dissertation provides a different applied example of an underutilized experimental design that can be used to evaluate programs for both pre- and in-service teachers.

One particularly nascent area of research on teacher professional development relates to educators' ability to use data to inform practice. As school systems have increased the amount and quality of the student data they collect through state and federal initiatives, teachers are expected to be able to review, analyze, and interpret this data to inform instruction (Coburn & Turner, 2011; Means, Padilla, & Gallagher, 2010). Additionally, the definition of data has moved beyond standardized test scores to incorporate all kinds of systematically collected information about students and their learning. In theory, incorporating this type of data to inform practice can bolster student learning as teachers are able to provide instruction targeted at students' demonstrated needs (Mandinach & Gummer, 2016). However, surveys of educational programs suggest that pre-service and in-service teacher training has yet to catch up with the demands to prepare teachers to use data in their classrooms (Gallagher, Means, & Padilla, 2008).

While some research has evaluated treatments intended to improve pre-service teachers' assessment and data literacy skills, there has been no experimental evidence on the effectiveness of these interventions. The first chapter seeks to fill the gap by providing causal research on data literacy supports within the context of a teacher

preparation program. In this experimental study, we evaluate a pre-service teacher workshop designed to improve participants' data literacy. We randomly assigned 90 pre-service teachers to attend a workshop and complete follow-up activities focused on enhancing their understanding of and comfort with different forms of student data. Rather than using a traditional parallel RCT in which one group serves as the untreated control group and the other receives treatment, we used an alternative-treatments design (Shadish et al., 2001). In this design, the comparison group receives a different substantive treatment focused on improving teacher candidates' cultural sensitivity. Relative to those who participated in the cultural sensitivity workshop, the data literacy intervention improved teacher candidates' attitudes about assessments, confidence in their data skills (.37 standard deviations, $p<.05$), and their longer-term instructional efficacy (.36 standard deviations, $p<.10$), providing causal support for the use of such interventions in teacher preparation programs.

Although the literature on data-use supports for *in-service* teachers is more developed relative to studies in pre-service settings (Carlson, Borman, & Robinson, 2011; Slavin, Cheung, Holmes, Madden, & Chamberlain, 2013), the research base is still small and largely descriptive (Ebbeler, Poortman, Schildkamp, & Pieters, 2017; Farley-Ripple & Buttram, 2014; Goertz, Olah, & Riggan, 2009). Additionally, none of these studies have examined data-use interventions intended specifically to support teachers as they interpret assessments administered as students enter school – a promising point for early, targeted intervention. As the use of statewide kindergarten entry assessments is growing across the country, it is important that we identify ways to help teachers on a large scale. In Chapter 2, I contribute to the existing evidence on effective data-use interventions by

evaluating the utility of data consultations to support kindergarten teachers' use of a statewide entry assessment. Additionally, I use a novel, two-stage randomization process that allows us to address questions about both the overall effectiveness of the data consultation in any form as well as the best ways to modify the intervention in order to make it more feasible at scale. To do this, I embed a factorial design varying two factors of the data consultations: format (one-on-one vs. group) and delivery (web-based vs. in-person). While we do not find that offering data consultations improved teachers' attitudes about data or data skills on average, our results suggest that the format and delivery of the intervention has implications for its effectiveness. Somewhat contrary to other research, we find that web-based, one-on-one interventions were most effective at improving teacher outcomes, both relative to any other format and to business as usual.

Chapter 3 contributes more explicitly to the methodological literature by outlining the advantages of repeated measures designs in educational settings and focusing on an underutilized variant: the experimental switching replication. This hybrid research design combines different three different design elements – repeated measures, randomization, and replication – in order to provide both causally and externally valid evidence of a program's impacts. The switching replication, therefore, provides a unique combination of internal and external validity within a single study when the two types of validity have traditionally been at odds with one another. This design is particularly well suited for educational contexts, as outcome observations are typically made multiple times over the course of the academic year and all study participants are able to receive treatment However, the experimental switching replication is poorly represented in both the methodological and applied educational literature, leaving researchers with little guidance

on how to implement such studies. This paper seeks to fill this gap by outlining the features of these different repeated measures approaches, and providing an example of an experimental switching replication within a teacher preparation context. We highlight the strengths of the design as it was used to evaluate coaching supports for teacher candidates as well as the challenges we encountered during implementation.

References

Carlson, D., Borman, G. D., & Robinson, M. (2011). A mulitstate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*, 378-398.

Coburn, C. E. & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement, 9*, 173-206. doi: 10.1080/15366367.2011.626729

Coburn, C. E. (2003). Rethinking scale: Moving beyond numbers to deep and lasting change. *Educational Researcher*, *32*(6), 3-12.

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2017). The effects of a data use intervention on educators' satisfaction and data literacy. *Education Assessment Evaluation Accountability, 29*, 83-105. doi: 10.1007/s11092-016-9251-z

Edmonds, W. A., & Kennedy, T. D. (2017). *An applied guide to research designs: Quantitative, qualitative, and mixed methods*. Los Angeles, CA: SAGE Publications

Farley-Ripple, E. N. & Buttram, J. L. (2014). Developing collaborative data use through professional learning communities: Early lessons from Delaware. *Studies in Educational Evaluation, 42*, 41-53.

Gallagher, L., Means, B., & Padilla, C. (2008). *Teachers' use of student data systems to improve instruction: 2005 to 2007*, ED-04-CO-0040/0002. Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., . . . Wei, T. E. (2016). *Focusing on mathematical knowledge: The impact of content-intensive teacher professional development* (NCEE 2016-4010). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Hamilton, L., Halverson, R., Jackson, S. S., Mandinach, E., & Supovitz, J. A. (2009). *Using student achievement data to support instructional decision making*. Washington, D.C.: U.S. Department of Education. Retrieved from http://repository.upenn.edu/gse_pubs/279

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95,* 798–812. doi:10.1016/j.jpubeco.2010.11.009

Jacob, A., & McGovern, K. (2015). *The mirage: Confronting the hard truth about our quest for teacher development*. Brooklyn, NY: TNTP. Retrieved from https://eric.ed.gov/?id=ED558206

Kraft, M.A., Blazer, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(1). doi: 10.3102/0034654318759268

Mandinach, E. B. & Gummer, E. S. (2016). *Data literacy for educators: Making it count in teacher preparation and Ppractice*. Teachers College Press. Kindle Edition.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of Education Data at the Local Level: From Accountability to Instructional Improvement*. Washington, DC: U.S.

Department of Education, Office of Planning, Evaluation, and Policy Development.

Scher, L., & O'Reilly, F. (2009). Professional development for K–12 math and science teachers: What do we really know? *Journal of Research on Educational Effectiveness*, 2. doi:10.1080/19345740802641527

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

Slavin, R.E., Cheung, A., Holmes, G.C., Madden, N.A., & Chamberlain, A. (2012). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, 50, 371-396. doi: 10.3102/0002831212466909

Stockard, J., & Wood, T. W. (2016). The threshold and inclusive approaches to determining "Best Available Evidence": An empirical analysis. *American Journal of Evaluation.* doi: 10.1177/1098214016662338

Developing Data Literacy: Investigating the Effects of Pre-Service Data Use Intervention

Katherine L. Miller-Bains, Vivian C. Wong, & Julie J. Cohen

University of Virginia

Study 1: Developing data literacy: Investigating the effects of pre-service data use intervention

Abstract

Currently, we have limited experimental evidence about how to support pre-service teachers' as they develop the skills, knowledge, and attitudes necessary to use of student data when they enter the classroom. In this randomized controlled trial, we assigned preservice teachers (n=90) to participate in a 2-hour workshop focused on data literacy, followed by a series of written reflection prompts in which they applied content in their semester-long teaching placements. Using the Conceptions of Assessment (Brown, 2006), the Teacher Self-Efficacy Scale (Tschannen-Moran & Hoy, 2001), and study-specific survey items, we measured participants' perceptions of assessments and data skills. Participants in the treatment condition reported significantly higher perceptions of assessment relevance, instructional self-efficacy, and data use skills relative to the control group.

Introduction

Over the past fifteen years, international and domestic policies have formalized expectations around the use of student data in educational settings (Deluca, LaPointe-McEwan, & Luhanga, 2016). Such policies have incentivized the creation of large-scale data systems (Institute of Education Sciences, 2015), increased requirements around data reporting (American Recovery and Reinvestment Act, 2009), improved the quality of information available to educators (National Center for Education Statistics, 2015), and promoted the use of evidence to inform practice (Council for the Accreditation of

Education Preparation, 2013). In the United States, changes to the ways in which educational entities collect, store, and organize student data have required substantial financial investments at the state and federal levels, and are motivated by the assumption that systematic use of data can lead to meaningful improvements in student learning. However, these investments have not necessitated changes to teachers' use of student data, in part because many teachers do not have the resources, comfort, or knowledge to engage in data-driven instruction (Farley-Ripple & Cho, 2014; Marsh, Pane, & Padilla, 2006; Means, Padilla, & Gallagher, 2010). This study examines one lever by which to encourage data-driven instructional practices by improving pre-service teachers' perceptions of data, confidence in their ability to use data, and knowledge of a variety of data.

While teachers have long used formative assessments to gauge student understanding and identify content to re-teach, these measures and their implementation have largely been informal in nature (Ingram, Mandinach, 2012; Schildkamp & Kuiper, 2010). Additionally, educators have relied heavily on intuition to guide instructional choices, sometimes deferring to prior experience rather than empirical data about student learning (Slavin, 2002). Although prior experience can serve as a useful guidepost, using it without multiple sources of information can lead to limited conclusions about students and learning (Ikemoto & Marsh, 2007). Alternatively, regularly incorporating broad evidence of student learning – a practice referred to as data-driven instruction – can allow teachers to draw more accurate inferences about students' strengths and needs (Coburn & Turner, 2011; Mandinach & Gummer, 2016).

Data-driven instruction differs from other models of teaching practice in multiple ways. First, it emphasizes intentional and ongoing use of diverse data (Datnow & Hubbard, 2016). Second, data-driven instruction broadens the definition of assessments to include *any* information systematically collected on students – from standardized tests to classroom observations (Boudett, City, & Murnane, 2013). Finally, teachers use different types of complementary data sources to gain a more complete understanding of student learning (Boudett, City, & Murnane, 2013; Mandinach & Gummer, 2016). In turn, teachers are able to better tailor and target interventions, and subsequently monitor progress (Means, Chen, & DeBarger, 2011).

While the literature base is small, multiple empirical studies suggest that this kind of ongoing data use can change teacher practices in ways that help to bolster student learning gains (Carlson et al., 2011; Konstantopoulos et al., 2013; Lai & McNaughton, 2016; van Geel, Keuning, Visscher, Fox, & 2016; Wiliam, Lee, Harrison, & Black, 2004). However, prior research finds many teachers lack understanding of and comfort with data, which in turn limits their implementation of data use in the classroom (Means et al, 2011; Piro, Dunlap, & Shutt, 2014; Wayman & Jimerson, 2014; Volante & Fazio, 2007). While many districts and states have instituted professional learning opportunities targeting in-service teachers' data use, these supports are often piecemeal and sporadic (Means, Padilla, & Gallagher, 2010) leaving teachers with widely varying abilities to engage with data to make instructional decisions (Farley-Ripple & Buttram, 2014; Goertz, Oláh, & Riggan, 2009; Ikemoto & Marsh, 2007; Means, Padilla, & Gallagher, 2010).

The need to equip all teachers with data use skills has led various groups to focus on teacher education programs as an avenue through which future educators can acquire these skills before entering the classroom (Mandinach, Friedman, & Gummer, 2015). However, survey research suggests that preparation programs have yet to catch up with the shifting demands on teachers to be data literate. Recent surveys of teacher preparation programs found that few incorporate courses or other structured opportunities for teacher candidates to learn about and engage with student data (DeLuca & Bellara, 2013; Greenberg & Walsh, 2012; Mandinach, Friedman, & Gummer, 2015). Additionally, there is little empirical evidence of the best ways in which to incorporate data education into teacher preparation (Reeves, 2017), and the limited research that is available regarding pre-service teachers' data use has largely been observational or descriptive in nature (Reeves, 2017). The lack of rigorous empirical research in the teacher preparation context further obscures the best ways in which to affect teacher candidates' data literacy.

In the present study, we evaluate an intervention targeting teacher candidates' knowledge and attitudes about educational data. To our knowledge, this study provides the first experimental evaluation of data literacy education implemented in a teacher preparation program. The treatment consisted of a brief workshop and a series of reflection prompts to help teacher candidates integrate concepts from the workshop into their semester-long student teaching internship. Because the workshop was short and low-cost, it was relatively easy to incorporate into the existing program curriculum. We find that pre-service teachers who participated in the treatment reported significantly higher perceptions of assessments, improved data skills, and long-term improvements in instructional self-efficacy relative to the control group. This intervention suggests that

teacher preparation programs provide a promising environment through which to develop future educators' attitudes and proficiency with data, and that such opportunities can be provided without significant alteration to the typical teacher preparation curriculum.

Data Use and Literacy

Educators are increasingly expected to use evidence rather than intuition to support instructional decisions (Ikemoto & Marsh, 2008; Slavin, 2002). Through a more data-driven instructional process, teachers refrain from making assumptions based solely on prior experience, and instead form inferences using evidence. Within this framework, data and evidence constitutes any information systematically collected on students and their learning, from classroom observations to standardized tests. However, in order to effectively incorporate evidence into their practice, teachers must be data literate (Gummer & Mandinach, 2015; Mandinach & Gummer, 2016). Data literacy represents the skills and knowledge necessary to access, understand, and evaluate different forms of student data (Althanses, Bennett, & Wahleithner, 2013; Means et al., 2011).

Educational researchers and professional communities have provided different definitions of the skills and attitudes encompassed in data literacy. Although these definitions have varied, they commonly conceptualize data literacy as an iterative and systematic process that includes developing hypotheses about student learning, recognizing the strengths and weaknesses of various sources of information, and the ability to combine multiple data types in ways that mitigate their individual weaknesses (Coburn & Turner, 2012; Mandinach & Gummer, 2016). In their descriptive study of educators data use practices, Ikemoto and Marsh (2008) suggest that these data use processes occur on a continuum of complexity. In the simplest form, data-driven

decision-making involves a single player (such as principal) using one high-level data source (such as a standardized test score) to make a decision. Educators more proficient in data use, on the other hand, generate hypotheses using multiple sources of data and use an iterative analytic process, drawing on additional information to probe alternative explanations or support the hypotheses. For example, Ikemoto and Marsh describe one exemplary school that engaged in more advanced data use. During this process, both the principals and the teachers used disaggregated data to discover that transfer students consistently scored higher on literacy exams. In consultation with literacy experts, the staff decided to adopt a curriculum used by other schools in the district. As Ikemoto and Marsh describe it: "Although they had hunches regarding the underlying causes of the data results, educators did not assume that these inclinations were correct. They combined evidence with expertise through a collective process to develop actionable knowledge and identify a solution" (p. 116).

Interventions Targeting Data Use

Previous research has found that data-driven instruction can lead to changes in teachers' classroom practices (Gearhart & Osmundson, 2009; Mertler, 2009) as well improvements in student learning (Carlson et al., 2011; Konstantopoulos et al., 2013; Lai & McNaughton, 2016; van Geel, Keuning, Visscher, Fox, & 2016). However, the evidence of data use's influence on student outcomes is mixed (Farrell & Marsh, 2016; Goertz, Olah, & Riggan, 2009). In one large-scale randomized controlled trial, Carlson and colleagues (2011) evaluated the effects of a district-wide, data-driven reform initiative on students' reading and mathematics achievement. Of the fifty eligible districts, half were randomly assigned to receive training in data use for school and

district administrators as well as the quarterly benchmark testing for students. After one year of implementation, the students in the treated districts scored significantly higher on state-administered math achievement tests by 0.06 standard deviations. A follow-up study (Slavin et al., 2013) found that the effects extended to reading achievement in the fourth year of implementation. Similarly, van Geel and colleagues (2016) found that a school-level data-based decision-making intervention improved students learning gains compared to growth in the two years prior to the intervention. As part of the intervention, teachers from 53 primary schools in the Netherlands engaged in team discussions and received coaching in data use from an external data expert over the course of two school years. The effects of the treatment were especially large in urban schools serving a larger proportion of low-income students.

Developing Data Literacy in the Teacher Workforce

With the advent of improved longitudinal data systems, teachers have access to more student data than ever before. However, surveys of in-service teachers suggest that educators are not always clear about how to use data effectively. Analyzing data from a large, nationally-representative study (National Educational Technology Trends Study), Gallagher, Means, and Padilla (2008) found that while the majority of teachers reported using data to inform parents, track student achievement, and monitor progress, less than half felt confident in their ability to interpret data. Additionally, more than one-quarter felt their professional development experiences had failed to prepare them for data use.

This lack of confidence is concerning given its negative relationship with the data use practices. Gallagher and colleagues found that low levels of confidence in data were significantly and negatively associated with th. Similarly, in their analysis of participants

from intensive data-use professional development, Dunn, Ariola, and Lo (2013) found that teachers' anxiety around data use was negatively associated with their sense of efficacy around implementing data use practices. Conversely, teachers who felt more confident in their ability to access and identify relevant student data as well as their ability to use available tools and technology were more likely to feel efficacious.

However, few teacher preparation programs have integrated data literacy concepts into their curriculum (DeLuca & Bellara, 2013; Greenberg & Walsh, 2012; Mandinach, Friedman, & Gummer, 2015). A recent review of 180 undergraduate and graduate schools of education by Greenberg and Walsh (2012) suggests that the vast majority of preparation programs do not provide teacher candidates with substantial opportunities to develop skills in three primary areas of data use: assessment literacy, data analysis, and linking data to instructional decisions (Greenberg & Walsh, 2012). More specifically, Greenberg and Walsh (2012) find that while many programs incorporated some course content related to assessments, only 10% of programs provided adequate opportunities to learn and practice data analysis and using data to inform instructional choices.

Furthermore, research suggests that the data use opportunities available to pre-service teachers leave them with limited perceptions of what constitutes data. This is partly because exposure to data-related content – both in clinical teaching and through coursework – tends to vary substantially from one pre-service teacher to the next depending on content area, student teaching placement, or grade-level. In a cross-sectional study, Volante and Fazio (2007) describe perceptions of data literacy in a sample of pre-services teachers at a Canadian school of education. The authors surveyed 69 teacher candidates in different stages of the same preparation program on their

perceptions and understanding of assessments, as well as suggested areas for further training. They found that, regardless of their year in the program, most respondents possessed a limited view of assessments as summative measures. Respondents also tended to favor personal communication and observations as more valuable sources of information about student learning, but did not specifically recognize these as student data. Furthermore, many of the preservice teachers did not feel confident in their ability to use data, regardless of far along they were in the program, and all suggested that they would benefit from additional training in assessment and evaluation.

Unlike Volante and Fazio, Athanases, Bennett, and Wahleithner (2013) found in their qualitative study of 80 preservice English language arts teachers the sample tended to draw from many different data sources when answering questions about student learning. These teacher candidates were able to pull from multiple data points – such as essays, surveys, and personal communications with students – in order to triangulate their findings and make better sense of student learning. This allowed them to engage in an inquiry process and develop important data literacy skills over the course of their clinical teaching experiences. However, many teacher candidates used only one or two data sources and did not collect further evidence to probe potential reasons for student performance, and were left to speculate about the root causes for patterns they initially observed.

Reeves (2017) also found significant variation in the opportunities afforded to individual teacher candidates across five different schools of education. However, he also suggested that clinical teaching provided teacher candidates with many chances to engage in data collection and use. Most teacher candidates reported having opportunities to use

data to modify instruction, determine students' levels of achievement, and evaluate the effectiveness of instruction at least once a week in their student teaching placements,. These experiences were enhanced by formal learning opportunities, as Reeves also found that those who had participated in teacher inquiry coursework were more likely to engage in data use in their student teaching. But enrollment in these formal courses was not consistent, as over one-third of teachers with no coursework in inquiry or teacher research.

Much like in-service teachers, pre-service teachers' data use practices are also influenced by their comfort with and exposure to data. In a survey study of Illinois public school teachers, Reeves, Summers, and Grove (2016) found that in-service teachers with exposure to courses around data literacy were more likely to use data in their practice. However, less than half of respondents reported having any taken classes related to inquiry or data use in as part of their undergraduate- or graduate-level coursework.

Empirical Studies on Data Literacy Interventions in Teacher Preparation

A handful of studies have evaluated data-use interventions implemented during teacher preparation programs. The results of this research suggests the potential for providing teacher candidates with data use experiences that broaden their perspective on data as well as enhancing their confidence and skills. In a series of qualitative and pre-post studies, Dunlap, Piro and colleagues evaluate the effects of a data literacy intervention embedded in a pre-service teacher preparation course (Dunlap & Piro, 2016; Piro & Hutchinson, 2014; Piro, Dunlap, & Shutt, 2014). The intervention, called Data Chat, introduced a multi-step process for using data, including analyzing data in teams, developing formative and summative assessments, and creating strategies based on

identified areas of need. Within the Data Chat, teams of pre-service teachers presented their data analyses and subsequent instructional plans. Based on participants' response to open-ended survey questions, the authors found that despite a general unease regarding data use prior to participation, the students sense of efficacy increased (Dunlap & Piro, 2016). Furthermore, a pre-post analysis suggested that the teacher candidates improved their ability analyze, interpret, and differentiate instruction based on student data, as most participants increased the percentage correct on the data skills assessment (Piro & Hutchinson, 2014). However, the authors do not provide significance tests and report the overall percentage correct across the entire group, making it difficult to determine the magnitude of the treatment effect.

Reeves and Honig (2015) investigated the effect of a 6-hour data course on teacher candidates' attitudes about data, as well as their data literacy skills. Using a pretest-posttest design, the authors collected information from 64 pre-service teachers attending a school of education. The intervention was provided within the context of a longer assessment course, during which the undergraduate teacher candidates learned to develop their own assessments. The authors measured outcomes using the Conceptions of Assessment (Brown, 2006), the Survey of Educators Data Use (Wayman, Cho, & Shaw, 2009), and researcher-developed measures of data literacy and overall satisfaction with the course. They found that participation in the course improved teacher candidates' data self-efficacy (0.23 standard deviations, $p<.01$), and perceptions the validity of assessments (0.31 standard deviations, $p<.05$) and their ability to hold students accountable relative to baseline (0.29 standard deviations, $p<.05$).

Current Study

The present study expands on recent examinations of the effects of data interventions in teacher educator programs and adds experimental evidence to the current literature base. Like Reeves and Honig (2015) and Piro et al. (2014), we look at participants' attitudes about assessments, confidence in their ability to use data, and general satisfaction after participating in short courses on data literacy. Our short course consisted of a two-hour workshop followed by a series of reflection prompts in which participants were asked to reflect on key concepts within their semester-long teaching internships. In moving from abstract knowledge to concrete application, we hoped participants would further develop their data literacy and skills in ways that would translate to their practicum and later classroom practice.While Reeves and Honig found evidence of the efficacy of the data intervention implemented within a similar sample, we incorporate a more rigorous research design by using a pre-post randomized experiment. This design allows us to isolate the causal effects of the data literacy intervention.

Methods

Participants

We worked with 90 students enrolled in a teacher preparation program at a large, selective public university in Southeastern United States. The teacher candidates were distributed across four different education programs: special education (19%), elementary education (48%), social studies education (22%), and English education (11%). All participants were enrolled in a one-credit seminar course that accompanied their semester-long teaching internships. The sample was predominantly white (85%) and female (78%). About half of the participants were in their fifth year of a Bachelor's/Master's program (57%), while the rest were enrolled in the second year of a

Post-Graduate Master's program (43%). Of those who initially consented to be in the study, 85 completed all intervention activities and had complete pre- and posttest data for both our proximal and long-term outcomes.

We present equivalence on pretreatment covariates across the two conditions as presented in Table 1. In order to assess whether the treatment and control conditions were balanced on key covariates, we estimate a multivariate regression including all baseline measures listed in Table 1 using Stata 14. The treatment and control groups did not display any significant baseline differences in terms of pretest scores, gender, grade point average, or proportion of white participants, as indicated by the individual t-tests as well as the test of joint significance. Additionally, we did not find that missing values differed across the treatment and control groups.

Research Design

Using a randomized experimental design, we examine the impact of treatment on participants' perceptions of assessments as well as the perceived utility of the intervention activities at the end of the academic semester. We also evaluate the longer-term effects on participants' instructional self-efficacy. Based on prior research, we anticipate that preservice teachers exposed to data literacy content would report higher levels of assessment relevance and utility and lower perceptions of assessment as a means of accountability.

To ensure an equal proportion of treatment and control participants across all programs and account for differences in faculty/instructional content, we employed a randomized block design including randomization strata for each course section. More specifically, we randomly assigned teacher candidates to treatment or control within each

of the four program areas with approximately 50% probability of assignment into either condition. Because teacher candidates within each program are exposed to slightly different content and classroom settings, we wanted to guarantee equal representation across the treatment conditions and account for any variation attributable to course section.

Current Intervention: Getting Data Wise

The current intervention focused on several aspects of pre-service teachers' data use (see Figure 1). Knowing that most teacher candidates feel ill-prepared to use data when they enter the classroom and lack exposure to data literacy concepts within their preparation programs, the intervention sought to improve participants' data literacy by improving their knowledge, skills, and perceptions. First, the intervention targeted teacher candidates' *knowledge* of assessments and data use practices through lecture, discussion, and small group activities. Specifically, the workshop content emphasized recognizing the strengths and weaknesses of different data sources, the use of multiple data sources in the process of triangulation, and making inferences based on evidence. During the workshop, teacher candidates were also given opportunities to practice *skills* by examining sample data, interpreting the evidence, and identifying other sources of information they would want in order to support their inferences.

Next, the intervention strengthened participants' data use *skills* and *perceptions* of assessments' usefulness through a series of brief, written prompts that had them apply the concepts using real-world classroom data. The participants completed the five writing prompts independently throughout the semester by collecting information and data from their semester-long student teaching internships (included in Appendix A). The prompts

All participants assigned to treatment completed the workshop and the prompts. In the long-term, we expected that their increased capacity and experiences with data use would improve their confidence and sense of efficacy around these practices in the classroom.

Data Literacy Workshop. The content of the workshop was adapted from Harvard University's online modules, *Introduction to Data Wise*. The open-access, 8-hour course is based on the Data Wise Project's work with schools and educators, summarized in the book *Data Wise: A step-by-step guide to using assessment results to improve teaching and learning* (Boudett, City, & Murnane, 2013). Boudett and colleagues designed Data Wise to provide teams of teachers with a framework through which they are able to engage with student data to make instructional decisions. As such, the course material does not focus on a particular grade or content area, and has been used by educators teaching in a variety of settings and subjects. Throughout the course, the authors provide data samples reflecting a variety of grade-levels and content. Using a combination of readings, videos, and activities, the course outlines the steps of the Data Wise process from organizing for collaborative work to acting on data.

For the present intervention, we concentrated on two of the Data Wise steps: (1) building assessment literacy and (2) digging into student data. We chose the material both because of time constraints and the applicability of the content to the teacher candidates' teaching internships. The authors of Data Wise designed the program to help build capacity within existing teams of in-service teachers and school or district administrators. As such, the later steps focus on using data collaboratively. While collaborative inquiry is an important component of data literacy, we could not guarantee that teacher candidates would be able to take part in such practices in their teaching placements. As such, we

chose to focus on the content that would support their habits of mind by providing them with protocols for examining data, a common vocabulary for understanding discussing data, and an understanding of how data can be used to support instructional decisions. Here, we emphasized the utility of data for making informed instructional decisions in the classroom rather than an external expectation imposed upon educators. We then incorporated elements from "digging into to student data" so that teacher candidates would have the chance to practice applying concepts to sample student data. Because we had teacher candidates from multiple programs working with students in various grade-levels, we did not modify the materials to be content or grade-level specific.

Follow-Up Reflection Prompts. Participants later applied strategies and concepts to their clinical teaching. These prompts focused on terms (e.g., validity, assessment differentiation) and data use practices (e.g., triangulation, noticing/wondering) covered during the workshop and had teacher candidates apply them to their teaching experiences. These prompts were intended to both reinforce the content of the workshop and allow teacher candidates to make connections to real-word practice. Each prompt consisted of three to four questions that required participants to examine different types of data collected in their classroom. They began by taking an inventory of all the assessments used in the classroom – from teacher-created assessments to statewide, standardized tests. They went on to think about potential gaps in the types of skills and learning domains that are covered by the different assessments, as well as things they would change in order to improve particular assessments. Finally, participants practiced combining the information from different assessments to identity a student learning need. They completed one prompt approximately every two weeks over three months, for a total of

five prompts. The prompts counted towards participation credit in the field experience course.

Control Condition. Teacher candidates in the control condition participated in a workshop of identical length and structure focused on cultivating relationships with students from different backgrounds. These participants also completed five prompts related to the content of the workshop over the course of the semester. Teacher candidates were given access to materials for their assigned treatment condition via a course repository. No one attended the workshop or completed prompts for the alternative study condition, and all participants turned in reflection prompts throughout the semester.

Measures

We collected information on participants' demographics, prior achievement, attitudes about data and assessments, and teaching self-efficacy via several different sources. We describe the specific outcome measures in more detail in subsequent sections. In order to collect the information, we administered online surveys to participants at two time points: immediately before the workshop at the start of the fall semester, and within one week of completing their final reflection prompts in late fall. We were able to obtain other relevant information, including demographics, background characteristics, and longer-term outcome data through the school of education's participant pool database. Teacher candidates complete measures and surveys at various points throughout their program as part of their research participation requirements. We used the information teacher candidates provided about their prior school experience, race, and gender in the survey completed upon entry into teacher education program.

Additionally, we utilize their responses collected at the end of their program – approximately 5 months after participants completed the intervention.

Proximal Outcomes. The Conceptions of Assessment (Brown, 2006) is a 27-item measure of educators' perceptions of the ways in which assessments can and/or should be used in schools and classrooms. More specifically, the instrument focuses on four constructs: assessment makes schools accountable, assessment makes students accountable, assessment improves education, and assessment is irrelevant. We list the statements included within each of the four subdomains in Table 2. All items use six, positively-packed selected response options (strongly agree, mostly disagree, slightly agree, moderately agree, mostly agree, and strongly agree). Items are averaged within each domain in order to produce an overall score for each construct. Prior research supports the COA-III's reliability and validity in large diverse samples of educators (Brown 2006, 2011). Three of the four subscales exhibit strong internal reliability in the current sample for both the pretest and the posttest (assessment makes schools accountable, assessment improves education, and assessment is irrelevant), while the fourth subscale displayed weak reliability at both administrations (assessment makes students accountable).

We also gave participants a brief survey at the conclusion of the study to gauge the helpfulness of the intervention in several targeted areas. In three Likert-scaled items, we asked whether the workshop and prompts helped teacher candidates to: (1) identify different uses of assessments in the classroom, (2) recognize the strengths and weaknesses of different types of assessments, or (3) combine different assessments.

Respondents indicated the utility of the workshop on a four-point scale ranging from not at all helpful to very helpful.

Long-Term Outcomes. The Teacher Self-Efficacy Scale (TSES; Tschannen-Moran et al., 2001) measures educators' sense of efficacy related to instructional strategies, classroom management, and student engagement (Table 3). Each item is rated on a 9-point scale with anchors at 1 ("Nothing"), 3 ("Very Little"), 5 ("Some Influence"), 7 ("Quite a bit"), and 9 ("A great deal"), and each subscale is made up of eight statements or questions. We use the instructional strategies subscale as our long-term outcome as respondents rate their ability to use a variety of assessment strategies and craft good questions for students. The TSES has been tested and used in many studies, and has demonstrated strong evidence of reliability and validity in many samples of teachers (Klassen et al., 2009; Tschannen-Moran & Woolfolk Hoy, 2001; Zee & Koomen, 2016). The reliability of the instructional self-efficacy subscale is strong in our sample at both pretest (α=.87) and posttest (α=.94).

Analytic Approach

To estimate the main treatment effects on our proximal outcomes, we run a series of regressions in Stata 14 using the following model:

$$Y_i = \beta_0 + \beta_1 Trt_i + \beta_2 Pretest_i + \boldsymbol{\delta}_1 Block_i + \varepsilon_i$$

We include a dummy indicator for treatment status (*Trt*) for each participant (*i*). This is our primary coefficient of interest, and indicates the difference on average between participants who participated in the workshop and those who did not on targeted

outcomes. We also include a control for the pretest measure when available ($Pretest$), and series of dummy indicators for the program blocks (elementary, special education, English, and social studies). We estimate robust standard errors. In the case of the individual Likert-scaled survey items, we also use the above regression model without pretest scores. While there has been some debate about the best way to analyze ordinal survey data as an outcome, simulation studies and sensitivity analyses suggest that regression analysis is an appropriate way to evaluate Likert-scale outcomes despite potential issues with non-normality (Norman, 2010; Sullivan & Artino, 2013). Therefore, we use the regression model to analyze our ordinal survey outcomes as well. Given our sample size the estimated explained variation from our randomization blocks and pretest covariates (r^2=.4 to .5), we will be able to reliably detect a treatment effect ranging from 0.40 – 0.46 standard deviations.

Results

We present the results of our outcome analysis in Table 4. The COA-III outcomes demonstrate that participation in the data literacy intervention substantially reduced perceptions that assessments are irrelevant to teaching (β=-0.37 sds, p<.05). However, we did not find significant improvements in participants' reports of the ability of assessments to improve teaching and learning as anticipated. Additionally, participation in the intervention did not affect teacher candidates' perceptions of assessments' ability to hold schools and students accountable.

We also asked participants about the usefulness of the workshop in three targeted areas: identifying use of different types of assessments, recognizing the strengths and weaknesses of different assessments, and combining information across data sources.

Across all three areas, teacher candidates reported significantly higher perceptions of the workshop's utility relative to the control group. On average, teacher candidates found that participation in the workshop was moderately helpful in these areas. In particular, participants reported much higher utility for recognizing the strengths and weaknesses of assessments, and generally found that the workshop was moderately useful in developing this skill. While the survey items also exhibit the extent to which the control condition provided opportunities to develop the skills targeted by the data literacy intervention, referred to as the treatment-control contrast, they also suggest that teacher candidates felt that the intervention helped them to develop skills above and beyond what was offered in the alternative workshop.

Examining the longer-term impacts of the intervention on participants' instructional self-efficacy four to five months after the intervention, we found substantial, marginally significant effects. Specifically, teacher candidates who completed the data literacy workshop and prompts rated their sense of instructional self-efficacy 0.36 standard deviations higher than their peers who participated in the alternative treatment (p=.07). Regressions on individual items comprising the self-efficacy subscale suggest that these effects were largely driven by increases in treated participants' comfort with using a variety of assessment strategies (β=0.38 sds, p=.06) and implementing different strategies in the classroom (β=0.53 sds, p=.06).

Discussion

National education organizations now expect teachers to use data and evidence to inform instructional practices. Incorporating data literacy content into teacher education programs presents a longer-term solution to developing human capacity than providing

professional learning opportunities to in-service teachers (Mandinach & Gummer, 2013), as schools of education are able to impart pre-service teachers with skills as well as shape teacher candidates attitudes and perceptions of assessments. These knowledge, perceptions, and attitudes serve as a precursor to data use, ultimately influencing educators' proclivity to engage in these practices to inform instruction (Datnow & Hubbard, 2015; Dunn, Ariola, & Lo, 2013; Dunn, Ariola, & Garrison, 2013). Although teachers often receive additional training in assessment use after they enter the field, these supports often leave in-service educators with gaps in their knowledge and understanding of how to use data to inform instructional practice (Means, Chen, Debarger, & Padilla, 2011; Means, Padilla, & Gallagher, 2010). Such gaps are further exacerbated by a lack of confidence and comfort in using data to inform instructional decisions, which make teachers less likely to try to use data to inform their practice (Datnow & Hubbard, 2015; Dunn, Ariola, & Lo, 2013).

Schools of education have seemingly struggled to find ways to incorporate opportunities for their teacher candidates to improve their attitudes around data use and acquire data skills through their coursework or teaching placements (Deluca & Klinger, 2010; Greenberg & Walsh, 2012; Mandinach, Friedman, & Gummer, 2015). The provision of data literacy coursework has been further complicated by a lack of guidance concerning the best ways to structure data literacy in undergraduate and graduate teacher education programs (Reeves, 2017). As a result, teacher candidates leave their preparation programs lacking the capacity to understand and interpret the various forms of student data despite the widespread expectations that they will use data to inform their practice (Athanses, Bennett, & Wahleithner, 2013; Reeves, 2012).

The findings of this study, however, suggest that a relatively low-intensity and low-cost intervention can improve teacher candidates' attitudes and skills with data, and that these improvements do not fade several months after participating. Other teacher preparation programs can easily incorporate the practices used in this study into their existing coursework using little time or resources, as the workshop required only two hours of facilitated class time and brief, ongoing reinforcement through reflection prompts. The magnitude of these effects are similar to those reported in previous observational studies with a self-selected sample and an intervention of somewhat higher intensity (Reeves, 2017). However, some of the expected areas of impact – specifically attitudes concerning the ability of assessments to improve teaching and learning – were not significantly altered by this intervention. Still, the effects of the intervention translated into increased perceptions of instructional efficacy as reported several months after the conclusion of the study, which indicate the potential longer-term benefits of such an intervention. As previous studies have found that improved instructional self-efficacy is associated with lower levels of teacher burnout (Fives, Hamman, & Olivarez, 2007; Zee & Koomen, 2016) and greater improvements in student motivation and achievement (Mojavezi & Tamiz, 2012; Zee & Koomen, 2016), the modest impacts of this type of light-touch intervention could translate into meaningful shifts on important outcomes. If the improvements in instructional self-efficacy transfer to the classroom and teaching practices, the students of these future teachers stand to gain both in terms of their engagement and learning.

Future Directions and Limitations

While this study demonstrates the promise for improving teacher candidates' data literacy through a low-intensity intervention, there are several areas that warrant further investigation and potential improvement. First, though we attempted to pull data examples from different grades and subjects to increase engagement with the workshop materials, we considered that participants may have found the content more authentic if we were able to make sample data specific to the content and/or grade level of the teacher candidates. In future iterations of the intervention, we plan to divide participants into subgroups that will allow us to tailor the data to their concentrations. The intervention was also limited by the inability to link these changes to improvements in student outcomes. Additionally, while prior research links teacher self-efficacy to student motivation and achievement (Mojavezi & Tamiz, 2012; Zee & Koomen, 2016), it is unclear whether the current participants will sustain their improved instructional self-efficacy as they enter the workforce and how these increases in self-efficacy and data knowledge might impact student learning. More research needs to be done to see how these practices carry over to the classroom as teacher candidates enter the education workforce. Finally, due to the lack of an established measure of data literacy skills, we relied primarily on teacher candidates' self-reports rather than direct assessments. While researchers are coming to a consensus on the basic skills that comprise data literacy, we need to develop scales that allow us to measure these skills in populations of pre- and in-service teachers.

References

American Recovery and Reinvestment Act (2009). *Pub. L. No. 111-5, 123 Stat. 115* (2009). Retrieved from www.gpo.gov/ fdsys/ pkg/ PLAW-111pub5content-detail.html

Athanases, S.Z., Bennett, L.H., & Wahleithner, J.M. (2013). Fostering data literacy through preservice teacher inquiry in English language arts. *The Teacher Educator, 48*(1), 8-28.

Boudett, K. P., City, E. A., & Murnane, R. J. (2013). *Data wise: A step by step guide to using assessment results to improve teaching and learning*. Cambridge, MA: Harvard Education Press.

Brown, G. T. (2006). Teachers' conceptions of assessment: Validation of an abridged version. *Psychological Reports, 99,* 166-170. doi: 10.2466/pr0.99.1.166-170

Carlson, D., Borman, G. D., & Robinson, M. (2011). A mulitstate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*, 378-398.

Coburn, C. E. & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement, 9,* 173-206. doi: 10.1080/15366367.2011.626729

Coburn, C. E., & Turner, E. O. (2011). The practice of data use: An introduction. *American Journal of Education, 118*(2), 99-111.

Council for the Accreditation of Education Preparation (2013). *CAEP accreditation standards of educator preparation*. Washington, DC: Author.

Datnow, A. & Hubbard, L. (2015). Teachers' use of assessment data to inform instruction: Lessons from the past and prospects for the future. *Teachers College Record, 117*, 1-25.

Datnow, A. & Hubbard, L. (2016). Teacher capacity for and beliefs about data-driven decision making: A literature review of international research. *Journal of Educational Change, 17*, 7-28.

Deluca, C. & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education, 64*, 356 - 372.

Deluca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Education Assessment Evaluation Accountability, 28*, 251-272.

Dunlap, K. & Piro, J. S. (2016). Diving into data: Developing the capacity for data literacy in teacher education. *Cogent Education, 3*, 1-13. doi: 10.1080/2331186X.2015.1132526

Dunn, K. E., Airola, D. T., & Lo, W. (2013). Becoming data driven: The influence of teachers' sense of efficacy on concerns related to data-driven decision making. *The Journal of Experimental Education, 81*, 222-241. doi: 10.1080/00220973.2012.699899

Farley-Ripple, E. N. & Buttram, J. L. (2014). Developing collaborative data use through professional learning communities: Early lessons from Delaware. *Studies in Educational Evaluation, 42*, 41-53.

Farrell, C.C., & Marsh, J.A. (2016). Contributing conditions: A qualitative comparative analysis of teachers' instructional responses to data. *Teaching and Teacher Education, 60,* 398-412.

Fives, H., Hamman, D., & Olivarez, A. (2007). Does burnout begin with student-teaching? Analyzing efficacy, burnout, and support during the student-teaching semester. *Teaching and Teacher Education, 23*(6), 916-934.

Gearhart, M., & Osmundson, E. (2009). Assessment portfolios as opportunities for teacher learning. *Educational Assessment, 14*(1), 1-24. ().

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *Can interim assessments be used for instructional change? Policy Brief.* RB-51. Phildelphia, PA: Consortium for Policy Research in Education. doi: 10.12698/cpre.2009.rb51

Greenberg, J. & Walsh, K. (2012). *What teacher preparation programs teach about K-12 assessment: A review.* National Council on Teacher Quality.

Gummer, E. S. & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record, 117*.

Ikemoto, G. S. & Marsh, J. A. (2007). Cutting through the data-driven mantra: Different conceptions of data-driven decision making. *Yearbook of the National Society for the Study of Education, 106*(1), 105-131.doi: https://doi.org/10.1111/j.1744-7984.2007.00099.x"

Ingram, D., Louis, S. K., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: barriers to the use of data to improve practice. *Teachers College Record, 106*(6).

Institute of Education Sciences (2015). *Draft statement of work: Regional Educational Laboratories 2011– 2016*. Washington, DC: Author.

Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis, 35*, 481-499.

Lai, M.K. & McNaughton, S. (2016). The impact of data use professional development on student achievement. *Teaching and Teacher Education, 60*, 434-443. doi: 10.1016/j.tate.2016.07.005

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*(2), 71-85.

Mandinach, E. B. & Gummer, E. S. (2016). *Data Literacy for Educators: Making It Count in Teacher Preparation and Practice*. New York, NY: Teachers College Press. Kindle Edition.

Mandinach, E. B., Friedman, J. M., & Gummer, E. S. (2015). How can schools of education help to build educators' capacity to use data? A systemic view of the issue. *Teachers College Record, 117*, 1-50.

Mandinach, E.B. & Gummer, E.S. (2013). A systematic view of implementing data literacy in educator preparation. *Educational Researcher, 42*, 30-37. doi: 10.3102/0013189X12459803

Marsh, J. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 14*(11).

Marsh, J. A., Pane, J. F., & Hamilton, L. S. (2006). *Making sense of data-driven decision making in education*. Santa Monica, CA: RAND Corporation.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' ability to use data to inform instruction: Challenges and supports*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Means, B., Padilla, C., & Gallagher, L. (2010). *Use of education data at the local level: From accountability to instructional improvement*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Metler, C.A. (2009). Teacher assessment knowledge and perceptions of the impact of classroom assessment professional development. *Improving Schools*, 12, 101-113. doi: 10.1177/1365480209105575

Mojavezi, A., & Tamiz, M. P. (2012). The impact of teacher self-efficacy on the students' motivation and achievement. *Theory and Practice in Language Studies, 2*(3), 483.

National Center for Education Statistics (2015). *Statewide longitudinal data systems grant program*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics. Retrieved from nces.ed.gov/ programs/ slds/ stateinfo.asp

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, 15(5), 625-632.

Piro, J.S. & Hutchinson, C.J. (2014). Using a Data Chat to teach instructional interventions: Student perceptions of data literacy in an assessment course. *The New Educator, 10*(2), 95-111. doi: 10.1080/1547688X.2014.898479

Piro, J.S., Dunlap, K., & Shutt, T. (2014). A collaborative Data Chat: Teaching summative assessment data use in pre-service teacher education. *Cogent Education, 1*. doi: 10.1080/2331186X.2014.968409

Reeves, T. D. (2017). Pre-service teachers' data use opportunities during student teaching. *Teaching and Teacher Education, 63*, 263-273.

Reeves, T. D., Summers, K. H., & Grove, E. (2016). Examining the landscape of teacher learning for data use: The case of Illinois. *Cogent Education, 3*. doi: 10.1080/2331186X.2016.1211476

Reeves, T.D. & Honig, S.L. (2015). A classroom data literacy intervention for pre-service teachers. *Teaching and Teacher Education, 50*, 90-101. doi: 10.1016/j.tate.2015.05.007

Schildkamp, K., & Kuiper, W. (2010). Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors. *Teaching and Teacher Education*, *26*(3).

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Sullivan, G. M., & Artino Jr, A. R. (2013). Analyzing and interpreting data from Likert-type scales. *Journal of Graduate Medical Education, 5*(4), 541-542.

Van Geel, M., Keuning, T., Visscher, A. J., & Fox, J. P. (2016). Assessing the effects of a school-wide data-based decision-making intervention on student achievement growth in primary schools. *American Educational Research Journal, 53*(2), 360-394.

Volante, L. & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implications for teacher education reform and professional development. *Canadian Journal of Education, 30*, 749-770.

Wayman, J. C. & Cho, V. (2009). Preparing educators to effectively use student data systems. In T. J. Kowalski, & T. J. Lasley (Eds.), *Handbook of data-based decision making in education* (pp. 89-104). New York: Routledge.

Wayman, J. C. & Jimerson, J. B. (2014). Teacher needs for data-related professional learning. *Studies in Educational Evaluation, 42*, 25-34.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: impact on student achievement. *Assessment in Education, 11*, 49-65. doi: 10.1080/0969594042000208994

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research, 86*(4), 981-1015.

Table 1. Baseline Differences between Treatment and Control Groups

	N	Constant	Mean Difference[a b]	P-Value
Conceptions of Assessment-III[c d]				
Assessment Makes Schools Accountable	90	0.01	-0.11	0.67
Assessment Makes Students Accountable	90	0.07	0.14	0.61
Assessment Improves Education	90	-0.10	0.03	0.90
Assessment Is Irrelevant	90	0.10	0.09	0.74
Teacher Self-Efficacy[d] – Instructional Strategies	83	0.11	0.10	0.74
Post-Graduate Masters of Teaching	78	0.47	-0.12	0.74
Grade Point Average [d]	65	0.07	-0.03	0.68
White	78	0.72	0.02	0.87
Female	77	0.74	-0.01	0.93
Overall F(13, 85)	*0.96*			

[a] Mean difference between students who participated in the data literacy workshops and subsequent prompts versus those who completed an alternative workshop and prompts.

[b] All estimates taken from a multivariate regression including all covariates as dependent variables with fixed effects for randomization blocks. The mean difference represents the slope coefficient associated with treatment indicator.

[c] Conceptions of Assessment –III Abridged

[d] All subscales standardized ($\frac{x-\bar{x}}{sd}$).

Table 2. Description of the Conceptions of Assessment-III (Brown, 2006)

Construct/Item	Reliability [a]	
	Pretest	Posttest
Assessment Makes Schools Accountable	.76	.83
Assessment provides information on how well schools are doing.		
Assessment is an accurate indicator of a school's quality		
Assessment is a good way to evaluate a school		
Assessment Makes Students Accountable	.32	.37
Assessment places students into categories		
Assessment is assigning a grade or level to student work		
Assessment determines if students meet qualification standards		
Assessment Improves Education	.90	.87
Assessment is a way to determine how much students have learned from teaching		
Assessment establishes what students have learned		
Assessment measures students' higher order thinking skills		
Assessment provides feedback to students about their performance		
Assessment feeds back to students their learning needs		
Assessment helps students improve their learning		
Assessment is integrated with teaching practice		
Assessment information modifies ongoing teaching of students		
Assessment allows different students to get different instruction		
Assessment results are trustworthy		
Assessment results are consistent		
Assessment results can be depended on		
Assessment is Irrelevant	.63	.75
Assessment forces teachers to teach in a way against their beliefs		
Assessment is unfair to students		
Assessment interferes with teaching		
Teachers conduct assessments but make little use of the results		
Assessment results are filed and ignored		
Assessment has little impact on teaching		
Assessment results should be treated cautiously given measurement error		

Teachers should take into account the error and imprecision in all assessment	
Assessment is an imprecise process	

[a] Reliability measured as internal reliability of each subscale at pretest and posttest.

Table 3.. Teacher Self-Efficacy Scale – Instructional Strategies (Tschannen-Moran & Hoy, 2001)

Item		
1) How well can you respond to difficult questions from your students?		
2) How much can you gauge student comprehension of what you have taught?		
3) To what extent can you craft good questions for your students?		
4) How much can you do to adjust your lessons to the proper level for individual students?		
5) How much can you use a variety of assessment strategies?		
6) To what extent can you provide an alternative explanation or example when students are confused?		
7) How well can you implement alternative strategies in your classroom?		
8) How well can you provide appropriate challenges for very capable students?		
	Pretest	*Posttest*
Overall Reliability[a]	*.87*	*.94*

[a] Reliability measured as internal reliability of each subscale at pretest and posttest.

Table 4. Outcomes from Data Literacy Workshop

	Treatment Mean[a] [SD]	Control Mean[a] [SD]	Beta Coef[b] (SE)
Conceptions of Assessment[c] (n=87)			
Assessment Makes Schools Accountable	-0.07 [1.00]	0.12 [0.98]	-0.05 (0.18)
Assessment Makes Students Accountable	-0.03 [1.07]	-0.07 [0.94]	-0.16 (0.19)
Assessment Improves Education	-0.06 [1.11]	0.09 [0.89]	0.02 (0.18)
Assessment Is Irrelevant[d]	-0.16 [0.94]	0.17 [1.06]	-0.37* (0.17)
Survey Items [e] (n=85)			
Identify different uses of assessments	2.84 [0.91]	2.29 [0.90]	0.54** (0.20)
Recognize strengths and weakness of different assessments	2.95 [0.94]	2.27 [0.90]	0.70*** (0.20)
Combine different data sources	2.70 [0.88]	2.13 [0.85]	0.56** (0.19)
Teaching Self-Efficacy Scale – Instructional Strategies (n=85)	0.16 [1.08]	-0.20 [0.88]	0.36^ (0.20)

Note: ^p<.10, *** p<.001, ** p<.01, * p<.05

[a] Simple treatment and control group means on the indicated variable. Standard deviations reported in brackets.

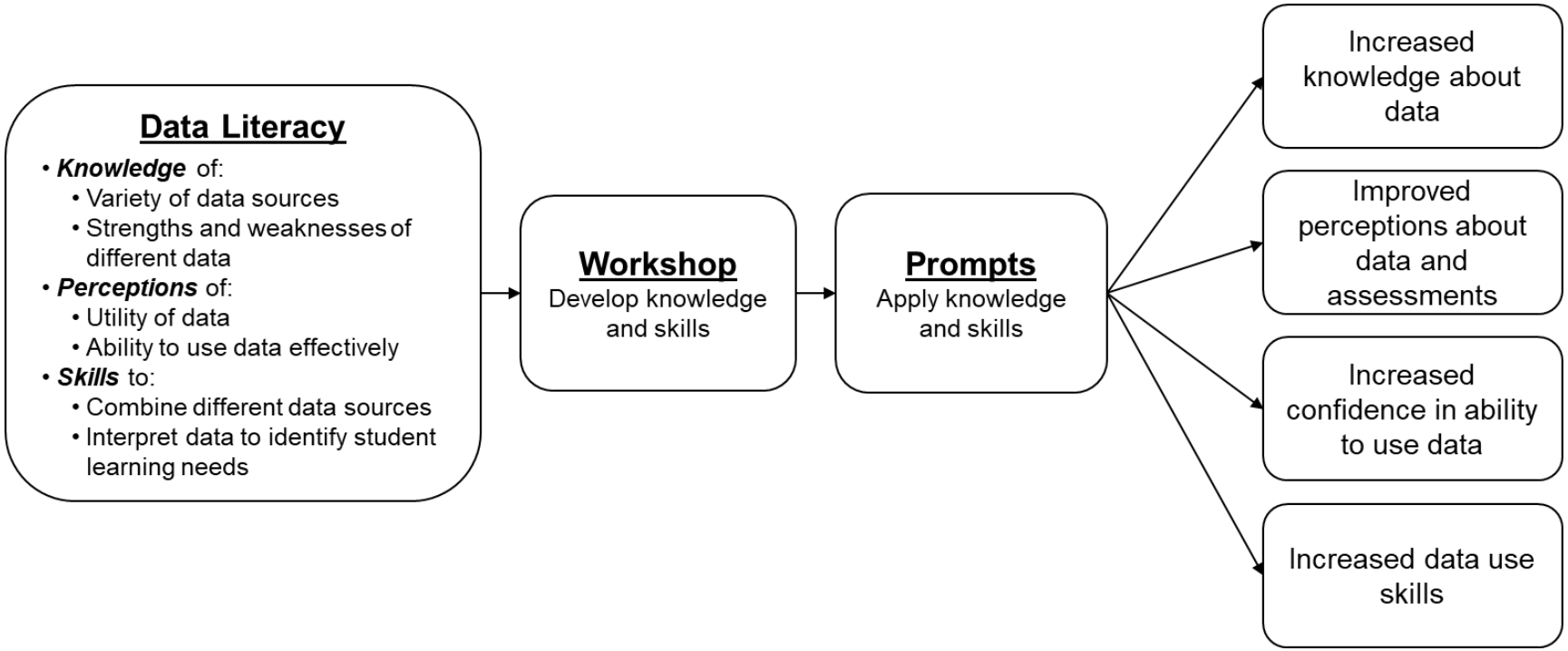
[b] All Conceptions of Assessment and Teaching Self-Efficacy subscale values have been standardized ($\frac{x-\bar{x}}{sd}$). Coefficients taken from separate regressions with indicated measure as the outcome and treatment assignment as the predictor. Each model contains fixed effects for randomization blocks. Regression of COA-III and TSES include controls for the pretest of the measure.

[c] Conceptions of Assessment Abridged, 3rd edition

[d] Higher scores indicate perception that assessment is less relevant to teaching and learning.

[e] All items rated on a 4-point Likert scale: Not at all helpful (1), Somewhat helpful (2), Moderately helpful (3), and Very helpful (4). Responses are not standardized.

Figure 1. Theory of Change



Appendix A
Data Literacy Writing Prompts

PROMPT 1:

1. Start by taking an inventory of the assessments used in your classroom. Make a list of the different measures that are used in your classroom (you can use the attached table to organize), being sure to include the following for each whenever possible/available:
    a. The learning domain assessed (e.g., math, literacy)
    b. The skills assessed (e.g., computation, analytical writing)
    c. The type of assessment (formative, summative)
    d. How/who created the measure
    e. How it is scored

   You may need to consult with your mentor teacher in order to compile the inventory.

2. Consider the coverage of skills and domains. Do you see any gaps in the skills assessed? Are there any assessments you would add?

3. Pick one assessment from the inventory. How reliable is the assessment? What would you change about the measure?

PROMPT 2:

Practice determining the quality/limitations of a particular assessment.

Pick another assessment from the inventory that **you feel could be improved**.

1. Discuss the characteristics of the assessment, being sure to address the following:
    a. How well does the assessment differentiate between students' skill levels?
    b. How confident are you in the assessment's consistency (e.g., reliability)?
    c. How well does the assessment cover the intended learning domain?
2. What would you do to improve the above assessment characteristics (differentiation, consistency, and coverage)?

PROMPT 3:

Practice looking closely at a single data source to identify a learner-centered problem. Please respond to following prompts.

a) Identify a question you have about student learning, and identify **one** assessment that will help you to answer that question.
b) Focusing on this data source, describe what this assessment can tell you about student thinking. Here it is helpful to look for patterns in students' responses. Are there

commonalities among items that students tend to get right?
c) Do any of the assessment results surprise you?
d) After considering the single data source, what other information would you want in order to address your original learner-centered problem?

PROMPT 4:

Practice using different sources of evidence.

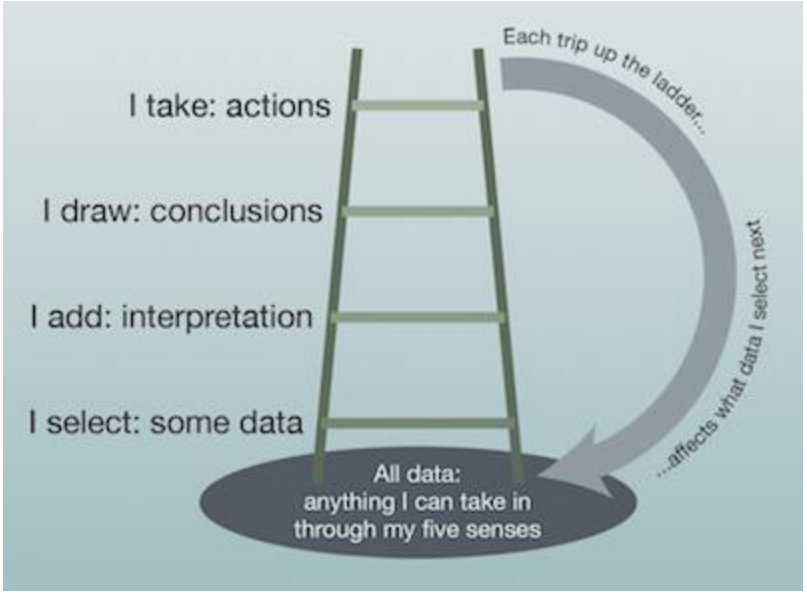
Identify one student and one skill area/learning domain for which you have multiple data sources (e.g., more than one assessment, observation, etc). Please answer the following:

a) Describe your impressions of the students' strengths based on each data source separately.
b) How do the results of each data source support or refute one another?
c) Describe what each source of information (the observations and the assessment) can and can't tell you about the students skills.
d) Do you feel the results point to particular areas of student need? Why or why not?

PROMPT 5:

First – update inventory as necessary with any additional assessments that have been conducted.

Choose a focus area, and practice making observations about the assessment results by moving up the ladder of inference. Collect the data that you have available related to performance in your chosen focus area.



a) Start with 3-5 things you **notice** about the assessment results. These should be facts with no interpretation.
b) List 3-5 things you **wonder** about the data. These should be framed as open-ended questions.
c) Next, brainstorm some hypotheses about the reasons for the patterns you observed above.
d) Consider what other evidence you would need to collect in order to confirm or refute the hypotheses you outlined above.

	Learning Domain	**Skills Assessed**	**Type of Assessment**	**Author of Assessment**	**How it is scored**
ex.	*Math*	*Addition*	*Summative (5-item quiz)*	*Ms. Baker*	*Percent Correct*
1					
2					
3					
4					

Digging into Data:

An Evaluation of Supports to Help Kindergarten Teachers Interpret Entry Assessments

Katherine L. Miller-Bains

Study 2: Digging into Data: An Evaluation of Supports to Help Kindergarten Teachers Interpret Entry Assessments

Abstract:

States across the country have instituted school entry assessments to provide kindergarten teachers with timely information about students' learning needs. However, teachers are often inadequately prepared to use this data to inform instruction. As educational offices introduce sweeping changes to the availability of entry data and expectations for their use, they must support teachers' data use practices on a large-scale. In this experimental study, we examined the implications for modifying data consultations designed to help teachers interpret entry assessments in order to allow for implementation across the state. We found that a more scalable form of the intervention, which used web conferencing to facilitate one-on-one data consultations between teachers and an external consultant, yielded the most consistently positive effects. This study concludes by highlighting the potential of incorporating experimental designs in order to develop more efficient, scalable interventions to support teachers as they work with kindergarten entry assessments.

Introduction

In 2011, the federal Race to the Top-Early Learning Challenge named statewide kindergarten entry assessments as a top priority for early education. Many departments of education have likewise invested in measuring students' skills upon kindergarten entry, with over 30 states piloting or implementing common entry assessments in 2017 (Center on Standards and Assessment Implementation, 2017). As entry assessments are administered within the first few weeks of school, they are intended to provide educators

with high-quality, quick information about students' incoming skills, allowing teachers to identify and target children's unique needs. Recognizing that school entry is a critical point to provide targeted supports (Belsky & MacKinnon, 1994; Chetty et al., 2011; Duncan et. al., 2007; Galindo & Sonnenschein, 2015; Hamre & Pianta, 2001), both policymakers and practitioners have promoted these measures as a means of decreasing achievement gaps in the short and long term (Meisels, 1998; RTT, 2011).

If entry assessments are to catalyze improvements in children's learning in the classroom, the results must be used in ways that lead to changes in teachers' practices and student learning. However, teachers' systematic use of large-scale assessment data has only recently become a professional expectation (Council for the Accreditation of Educator Preparation, 2013), and represents a departure for traditional modes of practice that rely more on heavily intuition and professional judgment (Ikemoto & Marsh, 2007; Ingram, Louis, & Schroder, 2004; Mandinach, 2012; Slavin, 2002). Given the growing use of statewide entry assessments and gaps in educators' data use skills, program developers must identify efficient and effective ways to support teachers' use of kindergarten entry assessments.

Traditionally, researchers approach the development and evaluation of such supports in a linear fashion – starting with observational studies in limited samples, then conducting efficacy evaluations of the developed program under ideal conditions, and finally implementing a randomized controlled field trial to evaluate effectiveness in larger contexts (Flay et al., 2005). However, this research process is often long and expensive (Bryk, 2015). Alternatively, some social scientists have adopted an iterative approach to developing and refining new programs, in which users interact with early

versions of an intervention within low-cost, randomized experiments. In particular, Collins and colleagues have suggested utilizing these screening experiments to investigate multiple intervention components at once using a factorial design (Collins, Dziak & Li, 2009; Collins, Murphy, Nair, & Strecher 2005; Collins et al., 2011; Collins, et al., 2007; Dziak et al., 2012). The factorial experiment can circumvent common issues associated with randomized-control trials, such as limited statistical power. Furthermore, the design permits direct evaluation of modifications to an intervention's features that allow for implementation at scale.

The current study uses a novel screening experiment in order to pilot and evaluate data use supports provided within a statewide entry assessment program. Specifically, we examined a consultation model in which teachers were able to review their classroom data with members of the research team, with the goal of identifying instructional actions related to their students' performance. We explored questions about the effectiveness and scalability of these supports implementing across a subset of diverse, randomly selected schools in the state. This permitted us to estimate the level of interest in these supports more broadly as well as challenges that should be addressed prior to statewide rollout.

Statewide Kindergarten Entry Assessments

Kindergarten entry assessments (KEAs) have been promoted as a promising means to improve student outcomes (Meisels, 1998; U.S. Department of Education, 2017). The emphasis on measuring skills at school entry stems from two lines of research on: (1) the importance of early learning for long-term success, and (2) the process of using student data to guide instructional decisions. This literature suggests that school readiness skills are important as they allow children to effectively engage in the

kindergarten classroom (Sabol & Pianta, 2012) and lay the foundation for future success (Claessens, Duncan, & Engel, 2009; Chetty et al., 2011; Duncan et al., 2007; Halle, Hair, Burchinal, Anderson, & Zaslow, 2012; Jimerson, Anderson, & Whipple, 2002; Miles & Stipek, 2006; Vitaro, Brendgen, Larose, & Trembaly, 2005). However, both natural variation in children's development and differential exposure to educational environments prior to school entry place students on unequal footing when they enter kindergarten. Without appropriate intervention, gaps in students' abilities tend to persist or grow as they progress through the school system (Pratt et al., 2016; Roy & Raver, 2014; Sektnan et al., 2010), underscoring the importance of identifying students' learning needs and providing them with appropriate supports as soon as possible.

In theory, large-scale kindergarten entry assessments provide educators with the timely and reliable information needed to provide students with targeted learning supports (Meisels, 1998; Snow, 2011). As a result, many states have instituted KEAs in order to inform policy and practice. While elementary schools have long used homegrown screeners or checklists to assess skills at kindergarten entry, these measures cover limited content and frequently do not have established psychometric properties (Shields et al., 2016). In contrast, the multi-dimensional, statewide entry assessments that have emerged in the past decade offer more comprehensive and reliable information about students (Little, Cohen-Vogel, & Curran, 2016; Miller-Bains et al., 2017). Furthermore, they provide classrooms, schools, and districts within the same state a common metric by which to examine the needs of incoming students. The use of these large-scale assessments has increased exponentially – from only seven states

implementing a common entry assessment in 2007 to more than 30 states in 2016 (Connors-Tadros, 2014; Weisenfeld, 2017).

Despite their growing popularity, research on the use of statewide entry assessment data as an instructional tool is sparse and mixed (Little et al., 2016; Shields et al., 2016). In one case study, Golan, Woodbridge, Davies-Mercer, and Pistorino (2016) examined the practices of kindergarten educators across four states in the early stages of implementing a kindergarten entry assessment. Through interviews and surveys, the authors found teachers and administrators did not use the results to make decisions about instruction or programs. Their use was further constrained by a lack of awareness of the information available to them or how to interpret the data reports. These respondents further suggested they would have benefited from explicit training on data use (Golan et al., 2016).

Using Student Data to Inform Instruction

Providing teachers with data through KEAs represents only one stage in a multi-step process, often referred to as data-driven instruction (see Figure 1). Through data-driven instruction, teachers collect and examine evidence in order to guide instructional decisions rather than relying primarily on intuition or prior experience (Mandinach, 2012). In this process, teachers first collect systematic information about their students through things like assessments, longitudinal databases, class assignments, or classroom observations. Then, they review the information in order to identify patterns in the data – such as a group of students scoring below benchmark on a particular skill. Teachers link these objective observations to needs that can be addressed in the classroom, and identify corresponding instructional actions or strategies including targeted lessons, activities, or

environmental modifications. After choosing an instructional action or strategy, they make a plan for implementation, including ways to monitor whether or not the chosen strategy is effective through future observations or assessments of students. They repeat this process as necessary, modifying supports based on students' progress.

When implemented successfully, data-driven instruction involves using disaggregated data to identify specific skills, examining things like subtest scores or exam content in order to determine discrete areas in which students need support in the classroom (Means, Padilla, & Gallagher, 2010). Instructional responses then focus specifically on supporting the identified skill(s). However, in their case study of thirty-six schools, Means and colleagues (2010) found that most schools tended to approach data use in a more perfunctory manner, reserving these practices to inform things like school-wide efforts to improve test scores rather than daily classroom-level instruction. Likewise, in their analysis of interview data collected from 18 Dutch primary teachers, Gelderblom and colleagues found that while most teachers reported using data to make instructional decisions, they usually made superficial use of the information (Gelderblom, Schildkamp, Pieters, & Ehren, 2016). For example, teachers might identify students that fell below average in a learning domain, but would not examine specific skill areas associated with the students' performance. They also failed to make instructional changes aligned with identified needs, seemingly re-teaching the same content rather than adapting their approach or pinpointing skills within the broader content area.

National- and district-level surveys further suggest that teachers at all educational levels lack the knowledge, confidence, and time necessary to use data (Gallagher, Means, & Padilla, 2008; Means, Chen, DeBarger, & Padilla, 2011). Despite attempts to provide

professional learning experiences to support teachers' data use, these efforts are often irregular and disjointed (Jimerson & Wayman, 2015; Wayman & Jimerson, 2014). Teachers, then, feel unprepared to use data to inform their practice, with their attitudes about data as well as their data-use anxiety and self-efficacy predicting the likelihood that they engage in data use practices to inform their instruction (Dunn, Ariola, & Lo, 2013; Gallagher et al., 2008). However, teachers in districts and schools that provide dedicated time and supports to examine data are more likely to use student data to inform instruction, suggesting that centralized support for data use at schools could lead to increased data-driven practices among teachers (Jimerson & Wayman, 2015; Wayman & Jimerson, 2014).

Recent empirical studies suggest that intensive, ongoing data-use interventions can improve teachers' perceptions of assessments, skills with data, and instructional practice as well as student learning. However, this evidence is mixed and stems mostly from observational studies. In the most rigorous evaluation of a data use intervention, Carlson and colleagues' randomly assigned 59 districts across seven states to receive a comprehensive data use intervention designed by the Johns Hopkins Center for Data-Driven Reform. The program included benchmark assessments, teacher and administrator professional development, and regular data meetings facilitated by an external expert in data use. Carlson and colleagues found that engaging in this intervention led to small but significant improvements in elementary and secondary students' math (.059 SD, $p<.01$) after the first year of implementation, as well as substantial and significant effects on elementary and secondary students' math achievement at year four (.32 and .31 SD, respectively, $p<.05$; Slavin et al., 2013). While the effects of the intervention were

consistently positive, they were not consistently significant across grade-levels and subjects. Slavin and colleagues (2013) noted the wide-range of implementation fidelity across participating schools and districts likely led to heterogeneous treatment effects across sites. Similarly, Keuning, Van Geel, and Visscher (2017) noted variability across schools implementing an ongoing data use intervention. The intervention consisted of a series of training sessions targeting participants' data skills, followed by a series of school-level meetings focused on analyzing students' math achievement and setting goals for improvement. The authors found that schools with higher quality teachers, more positive attitudes towards data use, and an existing culture of data use tended to experience larger positive gains on student achievement.

Designing Professional Supports for Use at Scale

While the above research suggests that direct intervention can help teachers improve their data use skills, states using entry assessments face the added challenge of delivering these supports at scale. Although most experts agree that high-quality PD should be embedded in the work environment, individualized, engaging, focused on discrete skills, timely, and of sustained duration (Desimone & Garet, 2015; Hill, 2007; Scher & O'Reilly, 2009), programs that possess these critical elements can be expensive and difficult to implement with many teachers across diverse geographic areas. Studies further suggest the difficulty of scaling up high-quality professional development initiatives as previously effective PD programs have failed to elicit similar changes in teachers' practices or student learning when scaled up to serve more participants (Garet et al., 2008; Garet et al., 2011; Kraft et al., 2018; Keuning et al., 2017).

In order to offer PD to large, diverse, and geographically-diffuse group of teachers, the provider must often modify the program in ways that might make the supports less effective. For example, coaching and consultation have been identified as some of the most consistently effective forms of professional support, largely because of their inherent ability to satisfy the characteristics of high-quality PD (Kraft et al., 2018). However, providing individualized, one-on-one coaching to all teachers within a district or state would not be possible without changing some of the common features of the PD (see Table 1). For instance, most consultations are one-on-one between a teacher and a coach, allowing for individualized and discrete support. Alternatively, providing coaching to a team of teachers would decrease the number of sessions needed within a larger sample while also opening up the possibility for collaboration. Yet, this modification could lead to less teacher engagement, decrease the ability to provide teachers with tailored support, and fail in schools lacking collaborative team dynamics. Similarly, conducting these sessions in-person allows teachers to feel more engaged and also provides a sense of job-embeddedness. However, this would require travelling to many schools across a large geographic area. While offering these session via web-conference would provide greater flexibility, it would likely result in less engagement, a potential loss of rapport between consultant and teacher, as well as the added challenge and unpredictability of technology.

Our understanding of how to scale-up PD in strategic ways is further limited by the current linear approach to the design and evaluation (Bryk, 2015; Cator & Adams, 2013; Tripp & Bichelmeyer, 1990), which is not well-suited to evaluating the implications of making these modifications. As Cator and Adams (2013) describe, the

"most widely accepted model today for determining the impact of a learning resource or intervention" involves three basic steps (p. 4). First, a small, often under-powered study is conducted to explore the basic principles of an intervention. This is followed by a somewhat larger study to investigate the treatment's effectiveness under ideal conditions, and concludes with a large-scale randomized controlled trial to establish efficacy in a real-world context (Cator & Adams, 2013). Although treatments may undergo an initial phase of refinement under this research framework, many interventions are conceived a priori with little exposure to authentic settings or users prior to scaling up (Bryk, 2015; Collins, Chakraborty, Murphy, & Strecher, 2009; Penuel & Fishman, 2012).

Alternatively, a more design-based approach to the development of supports focuses on creating efficient tools that operate well within real-world settings (Brown & Wyatt, 2010; Bryk, 2015; Collins, Murphy, Nair, & Strecher, 2005). Under this framework, simplified representations of the final product – or prototypes – are introduced to potential users in situations where prediction is difficult due to complexity, previous examinations did not produce satisfactory results, or there isn't a lot of prior informative research (Baek et al., 2008). Furthermore, this type of prototyping can be coupled with experimental designs such as the factorial in order to investigate multiple treatment components without the loss of statistical power (Collins, Dziak, & Li, 2009).

The present study sought to investigate how modifying such professional supports for use at scale may alter the effectiveness of teacher consultations using a design-based approach to evaluation. In particular, we focused these consultations on supporting teachers' data use within a kindergarten entry assessment program. The treatment consisted of data consultations with members of the research team and kindergarten

teachers. Given our limited capacity, we varied the format (one-on-one vs. group) and delivery (in-person vs. web) of the intervention in order to make the supports feasible when implemented across the state. The consultations focused on alleviating common barriers to data use by providing participants with dedicated time, effective protocols to guide the data-use process, relevant instructional resources, and a clear plan of action. Through this study, we addressed the following research questions:

1. What proportion of schools offered treatment take it up and do school characteristics predict participation?
2. Do data consultations affect teacher outcomes?
3. Does the format and/or delivery of the data consultation matter?

Research Methods

In order to determine the relationship between the different types of data supports as well as their interactions with one another, we conducted a randomized experiment with an embedded factorial design. To do this, we employed a two-step assignment process (Figure 2; discussed in more detail below). All randomization occurred at the school level with schools blocked by the number of classrooms to ensure balanced treatment allocation among small (2-3 teachers), medium (4-5 teachers), and large (6 or more teachers) schools. This meant that we were unable to include any schools with only one participating kindergarten classroom (n=14), leading to an initial sample of 132 schools. As we clustered assignment at the school level, all teachers within a given school received the same type of treatment.

In the first step, all schools were randomly assigned to receive data consultations or to business as usual (BAU). This contrast allows us to answer our first research

question by estimating the effects of receiving any sort of data consultation compared with the typical supports provided within schools. Because of the limited capacity to provide data consultations, a slightly larger proportion of schools in the sample were assigned to the BAU condition.

In the second step of the assignment process, we randomized all schools in the data consultation condition using a two-by-two factorial design, which allowed us to address the questions about the effect of varying the format and delivery of the intervention. Within a complete factorial design, units are randomly assigned to the different levels of each factor separately, and all possible treatment combinations are represented. This allows the researcher to estimate the main effects of each factor as well as interactions between them. In this case, the two factors vary: (1) whether the teachers received the consultation in a one-on-one or group format (referred to as format), and (2) whether or not the session was delivered in-person or via web meeting (referred to as delivery).

Current Kindergarten Readiness Program

Like many across the country, the state in the present study implemented a multi-dimensional assessment to gain more information about students' skills at kindergarten entry. The state had previously instituted a literacy measure, and the current program expanded into other early learning domains including mathematics and social-emotional learning. The program was in its third year at the time of the study, and individual districts elected to take part in the entry assessment. As part of the program, teachers administered a direct assessment of students' math skills and a rating scale of children's social-emotional skills via an online system within the first four to six weeks of the

school year. All participating teachers received training on how to administer the assessments, and had access to an online application that housed interactive data reports and links to recommended instructional resources based on students' assessment results. We describe the participating teachers, trainings, assessments, online system, and resources in more detail below.

Study Sample. Approximately 149 schools located in 44 school districts participated in the program in the 2016-17 school year. Of the 149 schools, 134 had more than one kindergarten classroom and were eligible for the study (see Table 2 for descriptives by condition). These schools were situated in rural (54%), city (25%), and suburban (22%) districts. On average, 4% of kindergarten students enrolled in these schools were eligible for special education services and 50% received free- or reduced-price lunch. Most schools in the sample (61%) served predominantly white students. The number of kindergarten classrooms within a school ranged from 2 to 14, and participating schools enrolled an average of 90 kindergarten students. Twenty-nine of the 59 schools offered data consultations took up the treatment.

Of the 350 teachers that completed a survey during their pre-assessment training, the majority reported having a bachelor's degree, 40% had a master's degree or higher, and most teachers had additional endorsements or certifications. On average, participating teachers had 14 years of teaching experience. Eighty-nine percent of teachers had taught in the same school the previous year, and of those teachers, 63% reported meeting in teams to discuss student data more than once per month in the last school year.

Pre-Assessment Teacher Trainings. In order to prepare teachers to administer the entry assessments, the research team provided in-person and webinar trainings for each participating district. These trainings were scheduled throughout the summer and in the weeks leading up to the assessment window. Because of variation in academic calendars and other scheduled professional development, districts participated in the trainings between 1 week and 1.5 months prior to administering the entry assessments. Each session lasted approximately 2.5 hours, and focused largely on accessing the materials via the online application, the content of the assessments, and standardization of administration. Teachers were given the chance to login into the online system and practice administering assessment items. At the end of the session, trainers showed teachers how to access the online data reports summarizing students' results and resources available to them through the online portal.

Entry Assessments Completed by Teachers. After completing these trainings, teachers administered two different assessments in their classrooms: a direct measure of children's math skills and a rating scale of social-emotional skills. The Early Mathematics Assessment System (Ginsburg & Pappas, 2016) – referred to as EMAS – covered children's emerging skills in numeracy, geometry, spatial sense, and patterning. During administration of the measure, an adult introduced each task using various stimuli (e.g., picture cards of balls and animals) and manipulatives (e.g., counters, paper). The administrator followed scripted prompts and responses that include scaffolding at various points throughout the assessment, allowing the student to see examples or receive standardized feedback on their responses. The EMAS includes 37 items and took an average of 20 minutes per child to complete. Teachers entered student responses in real

time during the assessment. All scripts and prompts were programmed in the online application so that the teacher would only see relevant instructions based on the current item and a student's response.

The social-emotional assessment, called the Child Behavior Rating Scale (CBRS: Bronson, Goodson, Layzer, & Love, 1990), captured information about students' social skills and self-regulation in the classroom. After observing each student for several weeks in the classroom at the start of the school year, the teacher rated the frequency with which a student displayed a described behavior on 5-point scale (from 1 indicating "never" to 5 indicating "always"). This information was entered directly into the online system outside of classroom time. The CBRS includes 17 items and took between two and three minutes to complete for each student.

Post-Assessment Online Reports and Resources. As soon as teachers entered information into the online system, they were able to access a series of automated data reports summarizing students' results. The interactive reports presented individual student or classroom scores in multiple formats: across the three broad readiness domains (e.g., math, social skills, self-regulation), in subdomains (e.g., numeracy, computation), and item-level scores. Scores were color-coded to indicate whether or not the student or class was meeting development expectations as determined by experts in math and socioemotional skills.

The reports also provided personalized links to instructional resources based on students' scores. Recommended resources were linked to areas in which the student or class scored below or just at expectation. These resources often took the form of short activities or lessons and focused on strengthening specific skills such as counting or

cooperating with peers. All of the resources were evidence-based and did not require materials beyond what was provided to teachers as part of their assessment kit or would be found in a typical kindergarten classroom. While the reports linked to suggested resources, teachers were able to access the entire bank regardless of the assessment scores.

Intervention - Data Consultations

In the prior school year, members of the research team implemented an experimental pilot of data use supports within one school district (Hasbrouck, 2016). As part of the intervention, researchers met individually with teachers to discuss entry assessment results and link these results to the instructional strategies highlighted above. While teachers who participated in these data consultations demonstrated improved perceptions of assessments and data skills, scheduling individual meetings with all teachers was beyond the capacity of the research team and would become increasingly infeasible on a larger scale.

Several members of the research team conducted data consultations with teachers following the same basic procedures. Principals at schools assigned to receive the consultation were contacted via email and offered a 25- to 50-minute discussion facilitated by a trained data consultant. These meetings were facilitated by a member of the research team. The meeting occurred one time approximately 1-3 weeks after the schools' assessment completion window and during individual or group planning periods. We hoped that by offering the consultations during this timeframe, all teachers at a given school would have enough time to complete both the math and social-emotional

measures, and the results of the entry assessments would still be relevant and useful for informing instruction.

Ahead of the meeting, the consultant asked teachers to choose a single learning domain covered in the readiness assessment (e.g., math, social skills, or self-regulation) as the area of focus during the consultation. At the start of the conversation, the consultants introduced the "I notice/I wonder" protocol to examine the teachers' assessment data (Venables, 2011), which asks participants to move intentionally from observable facts ("noticings") to inferences ("wonderings") supported by observed student data. This protocol is intended to slow down the interpretation process, help teachers to be more cognizant of assumptions, and ensure that inferences are based in patterns they see in the data. Next, the consultant had teachers link their observations about the data to potential areas of student support for a single student, small group, or entire class. Finally, the consultant and teachers collaborated to identify potential instructional strategies based on the demonstrated need, drawing from the bank of recommended resources provided by the readiness program and/or other sources available to the teacher. The teachers then completed an action plan (modified from the School Reform Initiative, 2017) detailing: (1) the planned change, (2) the evidence supporting the change, (3) what steps they planned to take to initiate the intervention, (4) any needed supports resources, and (5) their means of monitoring progress. Consultants followed up with participants two weeks after the data meeting to see if any other questions had surfaced as the teacher has attempted to implement their chosen instructional strategy.

Outcome Measures

We collected outcomes measures from all participating teachers between two weeks and one month after the consultations via an online survey. The research team sent survey links to teachers through principals and/or district-level administrators. We sent two targeted follow-up reminders to any schools with less than a 50-percent response rate. The outcomes described below represent a subset of the overall survey, which included additional questions about satisfaction and experience in the program. Through the survey, we assessed teachers' perceptions of assessments and data use using subscales from two measures - the Conceptions of Assessment III Abridged (COA-III; Brown, 2007) and the Survey of Data Use (SDU; Jimerson, 2016)

Subset of *Conceptions of Assessment*. The COA-III consists of 27 statements that load onto four latent constructs: Assessment Makes Schools Accountable, Assessment Makes Students Accountable, Assessment Improves Education, and Assessment is Irrelevant. The measure uses a positively-packed response scale in which teachers rate their level of agreement with each statement as: (1) strongly disagree, (2) mostly disagree, (3) slightly agree, (4) moderately agree, (5) mostly agree, and (6) strongly agree. Prior research established the internal structure of the longer COA-III (Brown, 2004) and confirmed the same first order factors in the abridged version when administered to two samples of primary and secondary teachers (Brown, 2007). We use a subset of the Assessment Improves Education subscale as an outcome in the present study. Teachers rated their agreement with the following four statements: (1) Assessment helps students improve their learning; (2) Assessment is integrated with teaching practice; (3) Assessment information modifies ongoing teaching of students; (4)

Assessment allows different students to get different instruction. The subscale demonstrated strong reliability in the current sample (α=.94).

Subset of *Survey of Data Use*. The SDU (Jimerson, 2016) asks teachers' to reflect on their data use practices, both individually and collaboratively. The 36-item survey is divided into 8 subscales: confidence, effectiveness of data-related professional learning, construal of data, beneficence of data, data anxiety, culture of collaboration, data vision/rationale, and professional learning communities. Respondents indicate their level of agreement with each item using a 5-point scale (strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree) except for those statements comprising the professional learning community block, in which they reported the presence/absence or frequency of different practices. In the initial development of the SDU, the items were piloted in three samples of teachers, and subscales exhibited moderate (α=.62) to strong reliability (α=.94). We use a subset of 6 items from the confidence subscale and 3 items from the culture of collaboration subscale, both of which exhibited strong reliability (α=.84 and α=.85, respectively).

Data-Skills Measure. Additionally, we created a direct assessment of teachers' skills in data use. We based the assessment on the relevant data-use competencies described by Means, Chen, Debarger, and Padilla (2011): data location (e.g., find relevant data in a graph or other representation), data interpretation (e.g., consider the distribution of scores across a group), and data use (e.g., drill down into subscale data in order to differentiate instruction). To assess these competencies, we presented teachers with sample data reports identical to those available through the assessment website. We then asked teachers to identify relevant data points and link to instructional actions or

strategies in series of 11 items that used both and closed responses. Six items were scored as correct or incorrect, and the remaining five items were scored as incorrect, partially correct, and correct. The data use assessment demonstrated fair reliability in the current sample (α=.57).

Satisfaction with Assessments for Instruction. Finally, we assessed teachers' overall satisfaction with and confidence in the two different assessments and their related instructional resources. Teachers responded to six questions each about the math (EMAS) and CBRS (social skill and self-regulation) assessments, including whether not they felt (1) confident in the results of the measure, (2) it accurately captured students' abilities, (3) it provided them with a better understanding of students' skills, (4) it was worth administering, (5) they used the measure to guide instruction, and (6) they would recommend the measure to other educators. To generate their overall satisfaction with each measure, we averaged responses across the six items. Each item used a five-point Likert scale ranging from "strongly disagree" to "strongly agree." Both scales demonstrated strong reliability in the current sample of respondents (α=.85 and α=.86 for the EMAS and CBRS questions, respectively).

Analytic Approach

We analyze all outcomes within a generalized regression framework in Stata 14. Our primary research questions necessitate different analytic approaches, and we describe them each below.

First, we estimate the overall effects of receiving the offer of any data consultation using the following intent-to-treat analytic model:

$$Y_{iS} = \beta_0 + \beta_1 DataConsultation_S + \delta_1 Block_S + \gamma_1 Characteristics_S + \tau_1 Characteristics_{iS} + \varepsilon_{iS}$$

Here, *i* indexes individual teachers situated in *s* schools. The variable $DataConsultation_{iS}$ is a dummy indicator of whether or not teacher was in a school that offered any sort of data consultation. β_1 is the primary coefficient of interest, and indicates the average difference between teachers in schools that were offered some sort of data consultation versus those in schools assigned to business-as-usual, controlling for a vector of teacher-level covariates and school-level randomization blocks and covariates. We included baseline measures of class size, the percent of white students enrolled in kindergarten, the percent of students qualifying for free- or reduced-price lunch, the percent of students qualifying for special education services, an indicator for rurality, and accreditation status. All continuous covariates have been centered around their mean, so that β_0 represents the average outcome when all dichotomous variables are set to zero and all covariates are at their mean. All standard errors are clustered at the school-level to account for the nesting of teachers in schools as the unit of treatment assignment. Given conventional significance level and power (α=.05, β=.2), our treatment and control group sizes and randomization blocks that explain about 0.10 of the variance in the outcome, we will be able to reliable detect effects of approximately 0.2 standard deviations in these analyses.

In order to analyze the factorial in the second randomization, we constrain our sample to only schools offered data consultations. These results allow us to see how the levels of the different factors compare to one another. We then use the following analytic model:

$$Y_{Si} = \beta_0 + \beta_1 InPerson_S + \beta_2 OneOnOne_S + \beta_3 InPerson * OneOnOne_S + \delta_1 Block_S + \gamma_1 Characteristics_S + \varepsilon_{Si}$$

In this case, we use effect coding for each factor as well as the dichotomous covariates. Effect coding differs from traditional dummy-coding in that treatment assignment is indicated by 1/-1, in the case of a two-level factor, rather than 0/1. This allows for the estimation of main effects and interaction effects to be uncorrelated (Kruger, Trail, Dziak, & Collins, 2016). This is appealing within a factorial framework as it allows the researcher to determine the effect of one factor *ignoring* the levels of another, rather than holding them constant. β_0 then represents the unweighted grand mean when all covariates are at their mean. β_1 represents the average difference between in-person and web-based data consultations from the conditional grand mean – ignoring the levels of the format factor. In other words, this coefficient represents the mean difference between all participants who were offered an in-person consultation versus those were offered a web-based consultation. Likewise, β_2 indicates the average difference between the one-on-one and the grand mean, and β_3 indicates the effect of one factor depends on the other. In other words, whether the effect of being assigned to an in-person consultation depends on whether or not the teacher is receiving a one-on-one or web-based consultation. All standard errors are clustered at the school level. Similarly, we will be able to detect effects of approximately 0.20 standard deviations (calculated using cluster factorial design in FactorialPowerPlan; Dziak, Collins, & Wagner, 2013).

Lastly, we contrast each data consultation format and delivery against the business as usual condition using the entire sample and the following model:

$$Y_{Si} = \beta_0 + \beta_1 InPerson\_OneOnOne_S + \beta_2 Web\_OneOnOne_S + \beta_3 InPerson\_Group_S + \beta_4 Web\_Group_S + \delta_1 Block_S + \gamma_1 Characteristics_S + \varepsilon_{Si}$$

Here, we include dummy-coded indicators for each of the four data consultation conditions (e.g., In-Person, One-on-One; Web, One-on-One; In-Person, Group; Web, Group) with the business-as-usual schools left as the uncoded comparison group. Therefore, coefficients β_1 through β_4 represent the effect of being assigned to the indicated data consultation condition relative to business as usual. Our power to detect significant treatment effects decreases, as we are now comparing treatment conditions with approximately 15 schools to those assigned to all nonattrited business-as-usual schools, making our minimum detectable effect size approximately 0.42 standard deviations.

Empirical Checks and Implementation Issues

Treatment Implementation. Implementation of the consultation model as described above occurred with mixed fidelity per field notes. Because participation was voluntary and we provided teachers and school administrators with a fair amount of discretion over these sessions, divergences emerged in terms of the communicated intent of the consultations, the actual format and delivery varied, as well as the components of the protocol that were covered.

First, some teachers entered the consultations with different understandings of the sessions' intended purpose. While we contacted principals with the same introductory email to offer the consultation, the degree to which and how the purpose of the meeting was conveyed to teachers varied. While many arrived with the expectation of discussing the assessment results, some teachers arrived having only been told that they would be

meeting with members of the program staff with no additional information. Others thought the consultants were there to simply hear their feedback on the assessment system. Because of this, a larger portion of the session was spent orienting teachers and/or listening to their concerns. This left consultants with limited time to review data with teachers. Additionally, some teachers expressed fundamental concerns with the utility of the data, and were less interested in linking the results to instructional actions.

The confusion over the intent of the consultations also led to some treatment crossover. In our original design, the consultations presented four distinct contrasts – (1) consultations conducted with an individual teacher and member of the research team at the school, (2) those conducted with the entire kindergarten team at the school using one teachers' data, (3) those facilitated with an individual teacher and a consultant using a video-conferencing application and (4) those conducted via web-conference with the kindergarten team. When conducting the web conferences, the intended format was generally achieved. Alternatively, when meeting in teachers' classrooms, other school personnel, students, or staff frequently entered during the consultations. In at least three schools, what was intended to be an in-person, one-on-one session turned into a meeting with the entire kindergarten team.

Per our consultation protocol, teachers were intended to access and review real classroom data, identify patterns, and complete an action plan based on their initial observations drawing from the online resources. While the consultants were generally able to review data with teachers and/or show them how to access the instructional resources that were directly linked to their students' assessment results, these intervention components did not always occur. In a subset of the in-school sessions, we encountered

technical issues that prevented consultants from being able to review data or resources with teachers. When visiting schools, consultants were often taken to conference rooms or other auxiliary spaces and accessed the data reports and resources on laptops. Because these were not DOE-sponsored devices, the team was unable to connect to the internet due to heightened security in connecting to schools' wireless networks. When possible, we would print a selection of the available data reports if unable to view using the online application. Conversely, all teachers were able to connect successfully to the web meetings, and because consultants had the capability to share screens, they were able to walk teachers through the data reports as the conversation progressed.

Attrition. Because of our school-level randomization process, we must consider multiple types of attrition (WWC, 2015). Figure 2 displays both the treatment take-up by and response rates by condition. For all analyses, we define attrited cases as missing outcome data. Overall, roughly half of all teachers completed the outcome measures, with slightly more teachers from the control condition responding (55% vs. 47%). Among the 59 schools assigned to receive data consultations, responses rates were roughly the same for those who took up treatment versus those that did not (45% for untreated vs. 50% for treated). Of schools that took up data consultation, teacher-level response rates ranged from 41 to 57%. Response rates for teachers in schools that did not take up treatment ranged from 36% in schools assigned to web-based, one-on-one to 58% in schools assigned to web-based, group sessions.

We must also consider differential cluster (e.g., school) and participant-level attrition, presented in Table 4. Here, we see that most schools did not fully attrite from the sample, meaning that they had at least one teacher response. Slightly fewer schools

who received the offer of treatment fully attrited from the sample relative to schools that were not offered data consultations. The differential attrition rate at the cluster-level is within the acceptable bounds set by the What Works Clearinghouse (WWC, 2015). However, we have much larger proportion of teacher-level attrition, as slightly more than half of teachers completed the outcome measures. Additionally, a higher number of teachers in schools assigned to business-as-usual responded to the survey.

Logistic regressions of baseline school characteristics on attrition status indicated that schools with a higher proportion of students falling below the benchmark in self-regulation and disadvantaged students were less likely to respond to the survey ($p<.10$). Looking at teacher-level attrition, none of the classroom characteristics (percent of students below benchmark, class size, etc.) were significantly associated with whether or not the teacher responded to the survey.

Results

Treatment Take-Up

As mentioned previously, approximately half (n=29) of the schools assigned to receive data consultations took up treatment. While the proportion of schools that took up treatment varied by condition – ranging from 39% of schools receiving an offer of web-based, one-on-one consultation to 64% of schools in the in-person, one-on-one condition – the differences were not statistically significant ($\chi^2=1.85$, $p>.05$). Because treatment take-up was determined by the principal or another school leader, we look at differences in school characteristics among those that took up treatment and those that did not using logistic regression on school covariates (Table 5). All coefficients are reported as odds ratios, meaning that values greater than one indicate that the likelihood of taking up

treatment increases as covariate increases, while values less than one suggest that the likelihood of taking up treatment decreases as the covariate increases. We found, controlling for all other school characteristics, schools were marginally more likely to take up treatment as the percent of students falling below the benchmark in social skills increased. Conversely, schools that only completed the CBRS, those with higher proportions of students falling below the benchmark in numeracy, and those with higher proportions of students attending preschool were significantly less likely to take part in the data consultations. Rurality, average class size, average student age, and percent white students were not significantly related to treatment take-up.

Effect of the Data Consultation on Teacher Attitudes and Skills

All results in Table 6 represent intent-to-treat estimates of being offered treatment and are reported in standard deviations. While teachers in schools that were offered additional supports did report higher levels of confidence, collaboration, or views of assessments on average, these results were not statistically significant ($p>.05$). Additionally, our treatment indicator and included covariates only explain a small portion of the variation in teacher responses, with r-squared values ranging from 0.03 to 0.15.

Factorial Effects of Delivery and Format

Next, we examined whether treatment effects differed based on either the delivery or the format of the data consultation. Heterogeneous treatment effects across the different factors could obscure overall treatment effects attributable to the consultation. As reported in Table 7, we found that overall, accounting for the different factors explains a much larger proportion of the variance in our outcomes of interest with r-squared values ranging from 0.11 to 0.31. In terms of format, we found that teachers who

received some form of one-on-one consultation tended to report higher but not significant levels of confidence in data use, attitudes about assessments, and data skills. Teachers who received any sort group-based consultation tended to report a greater sense of collaboration relative to those who were offered a one-on-one session with a data consultant, though this result was not significant ($p>.05$).

Examining the effect of delivery, we found that teachers who were assigned to the in-person condition reported lower average satisfaction with both entry assessments, and significantly lower satisfaction with the CBRS (-0.19 sds, $p<.05$) relative to the conditional grand mean. They also tended to score lower on the data skills assessment, though these results were not significant (-0.10 sds, $p>.05$). Teachers in the in-person condition did not appear to differ in terms of their overall confidence, sense of collaboration, or perceptions of assessments.

When examining the interactions, we found that the effects associated with each factor largely did not depend on the level of the alternate factor (e.g., the effect of one-on-one does not differ by in-person or web-based consultations, $p>.05$). However, teachers' data skills did differ based on the combination of format and delivery they receive (-0.24 sds, $p<.05$). The interaction indicated that teachers who were in the one-on-one, web-based condition had substantially higher scores than those in the one-on-one, in-person or group, web consultations.

Effects of Each Data Consultation Condition vs. Business-As-Usual

Because of the substantial variation in the effects of the different data consultations, we performed an exploratory analysis to estimate the effect of each treatment condition relative to teachers in the business-as-usual schools (Table 8). This

analysis reveals that, overall, the web-based, one-on-one consultations tended to yield positive results across the outcomes relative to the control group. Specifically, the teachers in schools assigned to the web-based, one-on-one condition rated their data-use confidence (0.30 sds, $p<.10$) as well as their sense of collaboration (0.38 sds, $p<.05$) higher than teachers in control schools. While not rising to the level of statistical significance, teachers in the web-based, one-on-one consultations also tended to demonstrate higher levels of confidence in assessments and data skills than teachers in control schools ($p>.05$). Teachers in schools that received *in-person*, one-on-one consultations similarly scored higher in data use confidence (0.36 sds, $p<.10$) and perceptions of assessments (0.44 sds, $p<.05$). However, teachers in this condition tended to perform worse on the data skills assessment, though this result was not statistically significant (-0.35 sds, $p>.05$).

Discussion

As the use of kindergarten entry assessments grows, so do states' need to support teachers, schools, and districts in making use of the information produced by these measures. Yet, the current evidence base for both KEAs and data use interventions provides limited guidance on how to best support these processes. Developing effective and efficient supports for use at scale is further complicated by a research process that does not often yield timely results or interventions that easily adapted for use in larger, more diverse contexts. However, unconventional experimental methods such as the ones utilized in this study provide a potential means of simultaneously delivering and evaluating theses supports in a more iterative manner.

In the current study, the embedded factorial design provided us with valuable information about the potential ways to modify a previously piloted data use intervention for delivery at scale without sacrificing potency. While maintaining the low-cost of the study restricted the outcome measures we were able to collect, the overall design uncovered both issues with implementation as well as promising results for the scalability of these supports that did not emerge in the smaller pilot study. We found that although many teachers and division leaders had previously expressed interest in these types of data-use supports, only about half of schools offered data consultations took them up. Furthermore, it seems that interest differed by school characteristics, as schools with previously low academic performance and higher need students in terms of disability and English-learner status were more likely to take up treatment. Also, those schools in which teachers' rated a higher number of students as falling below the benchmark in social skills were more likely to take up treatment, while the opposite was true across most math domains. This might suggest that schools felt they needed more help supporting the development of social skills in the classroom as compared to math skills.

The consultations themselves yielded mixed results. Prior research suggests that teachers' knowledge of, attitudes towards, and confidence with data are interrelated (Dunn, Ariola, & Garrison, 2013). While intuition might suggest that as one of these factors increases, the others follow, the development of new skills and practices is punctuated by in. For instance, under the concerns-based adoption model (Hall & Hord, 1987; 2011), teachers often start as generally unconcerned about a new innovation, becoming reluctant and avoidant before either incorporating the new practice or exploring what they consider to be more viable alternatives. This means as teachers'

sense of confidence increases and they engage in more data use, they are more likely to express concerns about data-driven practices and dissatisfaction (Dunn et al., 2013; Dunn & Rakes, 2011). This was somewhat evident in the pattern of our results, as teachers who participated in the data consultations were less likely to be satisfied with the assessments themselves while expressing more confidence in data and their ability to use this information effectively. In order to provide appropriate professional supports, we must improve our understanding of why teachers find this process helpful and what alternatives they would prefer over data-driven instruction.

Additionally, expanding our sample beyond a single district in which teachers volunteered to participate led to key differences in the teachers and schools participating in the intervention. In practice, we found that teachers varied in their: (1) understanding of assessments, (2) willingness to engage in discussions around the data, (3) experience with the assessments and system, (4) schools' and divisions' regular use of data, and (5) openness to identifying instructional strategies or activities to use in their classrooms. These variations combined with the teacher-directed nature of these conversations and the wide-range of meeting times (between 20 minutes to one hour) meant that components of the data consultation meetings occurred irregularly and the wide-range of treatment effects from site to site. Differences across schools in terms of existing practices and attitudes were similarly observed in other large-scale studies of data-use interventions (Carlson et al., 2011; Lai & McNaughton, 2016; Slavin et al, 2013), which led to heterogeneous treatment effects. This suggests the need to provide differential data-use supports to teachers and schools depending on their baseline perceptions and skills.

Finally, using a factorial to explore the modification of the consultations for use at scale produced some of our results. While we anticipated that the web-based consultation sessions would pose the most technical and implementation challenges, these issues tended to be more prevalent in meetings with teachers on school campuses, where firewalls or nonsecure spaces prohibited consultants from pulling up teachers' data reports. Furthermore, although we anticipated that the in-person, one-on-one sessions would yield the largest treatment effects, our evidence suggests that the web-based, one-on-one sessions tended to result in the most consistently positive effects on teachers' attitudes and skills. We believe this is in part due to the implementation issues we experienced in schools, as technology delays and less control over the structure of the consultations sometimes prevented the research team and teachers from reviewing the data. On the other hand, consultants had full control over the screen and pace of the conversation when reviewing data with teachers in one-on-one web meetings. These findings provide a launching point for our continued development and evaluation of these data use supports.

Limitations and Future Directions

We should be cautious in our interpretation of the results of the study due to the large amount of missing data in the pretreatment and outcome measures as well as the limited take-up of the treatment, as these attrition rates are above thresholds established by the What Works Clearinghouse. Although we found few significant differences on the observed baseline characteristics and conducted intent-to-treat analyses, there may be unobserved and important differences between schools that decide to take-up treatment as

well as those teachers who responded to the survey measures. All of these factors can bias our observed treatment effects.

Future studies should attempt to connect administrative and analytic data in order to reduce the amount of missing data and alleviate the burden on participants to respond to multiple instruments. The research team is already working to disaggregate website analytic data for future years in order to identify how individual teachers engage with the online application and resources. Additionally, the present study does not speak to the effects of data use practices on student outcomes. At the time of implementation, we did not have a proximal measure of students' outcomes. However, future iterations of the readiness measure will include spring scores for students' allowing us to connect teacher practices to student outcomes within a school year. Given the general positive effects of web-based consultation, future studies should explore this as a potential means of delivering data supports using an external consultant.

Lastly, the study was limited by the lack of instruments measuring teachers' data skills. The marginal amount of outcome variance explained by our models, in which r^2 values ranged from .04 to .31, suggests that many other factors are associated with data attitudes and skills that are not captured by readily available covariates. More research needs to be done in order to understand what predicts teachers' proclivity and ability to use data to inform instruction, while additional measures should be established that efficiently capture this information.

Conclusions

Much of understanding of what works and what does when supporting teachers' professional needs is based in theory. While the most highly-regarded features of these

supports – such as job-embeddedness, sustained, and individualized – are desirable and would intuitively lead to larger impacts on teacher outcomes, they are not pragmatic in real-world school settings that must serve diverse and diffuse groups of teachers (Kraft et al., 2017). Additionally, the results of this study call into question the assumption that more intensive professional supports translate into larger effects above and beyond more scalable formats. Instead, it seems that maintaining some elements of high-quality professional development, like the individualization, while forgoing others, can be as or more effective than trying to implement all components simultaneously.

Furthermore, understanding how interventions will function on a large-scale can be hindered by the typical, linear approach to development. Instead, we should utilize randomized experiments in order to empirically determine the implications of modifying features of a program for implementation at scale. Because capacity is inherently restricted in these early stages of development, randomization is more politically feasible, as it ensures equal access to these limited supports. Additionally, certain experimental designs such as the factorial, can answer questions about how to most efficiently and effectively structure supports in the future. This means that the end product will only include those elements that are necessary, and researchers can directly evaluate the implications of modifying the supports for use at scale. Lastly, field testing supports on a broad scale illuminates challenges and responses that be difficult to anticipate without user interaction. Rather than piloting burgeoning interventions in small samples, programs can benefit from including as many contexts as possible in order to avoid issues that will arise when supports are offered to the larger program population. These kinds of methods are valuable within the kindergarten entry context, as states must meet the needs

of teachers while also scaling up assessment efforts. While the results were not conclusive, they provide key information and direction as we move forward with refining the intervention to serve diverse needs.

References

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations, 61*(8), 1139-1160.

Belsky, J., & MacKinnon, C. (1994). Transition to school: Developmental trajectories and school experiences. *Early Education and Development, 5*(2), 106-119.

Brown, G. T. (2006). Teachers' conceptions of assessment: Validation of an abridged version. *Psychological Reports, 99*, 166-170. doi: 10.2466/pr0.99.1.166-170

Bryk, A. S. (2015). 2014 AERA Distinguished Lecture Accelerating How We Learn to Improve. *Educational Researcher, 44*(9), 467-477.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A mulitstate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis, 33*, 378-398.

Center on Standards and Assessment Implementation. (2016). Overview of major assessment types in standards-based instruction. San Francisco, CA: WestEd. Retrieved from http://www.csai-online.org/sites/default/files/resources/6257/CSAI_AssessmentTypes.pdf

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. *The Quarterly Journal of Economics, 126*(4), 1593-1660.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*(4), 415-427.

Collins, L. M., Chakraborty, B., Murphy, S. A., & Strecher, V. (2009). Comparison of a phased experimental approach and a single randomized clinical trial for developing multicomponent behavioral interventions. *Clinical Trials, 6*(1), 5-15.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: a resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*(3), 202

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

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine, 30*(1), 65-73.

Connors-Tadros, L. (2014). *Information and resources on developing state policy on kindergarten entry assessment (KEA)* (CEELO FASTFacts). New Brunswick, NJ: Center on Enhancing Early Learning Outcomes. Retrieved from http://ceelo.org/wpcontent/uploads/2014/02/KEA_Fast_Fact_Feb_11_2014_2.pdf

Council for the Accreditation of Educator Preparation (2013a). Standard 2: Clinical partnerships and practice. Retrieved from http://caepnet.org/standards/standard-2.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428–1446.

Dunn, K. E., Airola, D. T., & Garrison, M. (2013). Concerns, knowledge, and efficacy: An application of the teacher change model to data driven decision-making professional development. *Creative Education*, *4*(10).

Dziak, J. J., Nahum-Shani, I., & Collins, L. M. (2012). Multilevel factorial experiments for developing behavioral interventions: power, sample size, and resource considerations. *Psychological Methods, 17*(2), 153.

Dziak, J. J., Collins, L. M. & Wagner, A. T. (2013). *FactorialPowerPlan users' guide (Version 1.0)*. University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu ().

Ebbeler, J., Poortman, C. L., Schildkamp, K., & Pieters, J. M. (2017). The effects of a data use intervention on educators' satisfaction and data literacy. *Education Assessment Evaluation Accountability, 29*, 83-105. doi: 10.1007/s11092-016-9251-z

Farley-Ripple, E. N. & Buttram, J. L. (2014). Developing collaborative data use through professional learning communities: Early lessons from Delaware. *Studies in Educational Evaluation, 42*, 41-53.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., ... & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention science*, *6*(3), 151-175.

Galindo, C., & Sonnenschein, S. (2015). Decreasing the SES math achievement gap: Initial math proficiency and home learning environments. *Contemporary Educational Psychology*, *43*, 25-38.

Gallagher, L., Means, B., & Padilla, C. (2008). *Teachers' use of student data systems to improve instruction: 2005 to 2007*, ED-04-CO-0040/0002. Washington, D.C.: U.S. Department of Education, Office of Planning, Evaluation and Policy Development.

Golan, S., Woodbridge, M., Davies-Mercier, B., & Pistorino, C. (2016). *Case studies of the early implementation of kindergarten entry assessments*. Washington, D.C.: U.S. Department of Education.

Hall & Hord, 1987; 2011

Halle, T. G., Hair, E. C., Burchinal, M., Anderson, R., & Zaslow, M. (2012). *In the running for successful outcomes: Exploring the evidence for thresholds of school readiness*. US Department of Health and Human Services

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child development*, *72*(2), 625-638.

Ikemoto, G. S., & Marsh, J. A. (2007). chapter 5 Cutting Through the "Data-Driven" Mantra: Different Conceptions of Data-Driven Decision Making. *Yearbook of the National Society for the Study of Education*, *106*(1), 105-131.

Ingram, D., Louis, K. S., & Schroeder, R. G. (2004). Accountability policies and teacher decision making: Barriers to the use of data to improve practice. *Teachers College Record*, *106*(6), 1258-1287.

Jimerson, S. R., Anderson, G. E., & Whipple, A. D. (2002). Winning the battle and losing the war: Examining the relation between grade retention and dropping out of high school. *Psychology in the Schools*, *39*(4), 441-457.

Keuning, T., Van Geel, M., & Visscher, A. (2017). Why a data-based decision-making intervention works in some schools and not others. *Learning Disabilities Research & Practice*, 1-14. doi: 10.1111/ldrp.12124

Lai, M.K. & McNaughton, S. (2016). The impact of data use professional development on student achievement. Teaching and Teacher Education, *60*, 434-443. doi: 10.1016/j.tate.2016.07.005

Little, Cohen-Vogel, & Curran, 2016Little, M. H., Cohen-Vogel, L., & Curran, F. C. (2016). Facilitating the transition to kindergarten: What ECLS-K data tell us about school practices then and now. *AERA Open, 2*(3), doi:2332858416655766.

Mandinach, E. B. & Gummer, E. S. (2016). *Data Literacy for Educators: Making It Count in Teacher Preparation and Practice*. Teachers College Press. Kindle Edition.

Means, B., Chen, E., DeBarger, A., & Padilla, C. (2011). *Teachers' Ability to Use Data to Inform Instruction: Challenges and Supports*. Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development.

Meisels (1998). *Assessing Readiness.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

Melter, C.A. (2009). Teacher assessment knowledge and perceptions of the impact of classroom assessment professional development. *Improving Schools*, *12*, 101-113. doi: 10.1177/1365480209105575

Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development, 77*(1), 103-117.

Pratt, M. E., McClelland, M. M., Swanson, J., & Lipscomb, S. T. (2016). Family risk profiles and school readiness: A person-centered approach. *Early Childhood Research Quarterly, 36*, 462–474. doi:10.1016/j.ecresq.2016.01.017.

Reynolds, A. J., & Temple, J. A. (2008). Cost-effective early childhood development programs from preschool to third grade. *Annu. Rev. Clin. Psychol., 4*, 109-139.

Roy, A. L., & Raver, C. C. (2014). Are all risks equal? Early experiences of poverty-related risk and children's functioning. *Journal of Family Psychology, 28*(3), 391.

Sabol, T. J., & Pianta, R. C. (2012). Patterns of school readiness forecast achievement and socioemotional development at the end of elementary school. *Child Development, 83*(1), 282-299.

Sektnan, M., McClelland, M. M., Acock, A., & Morrison, F. J. (2010). Relations between early family risk, children's behavioral regulation, and academic achievement. *Early Childhood Research Quarterly, 25*(4), 464-479. ().

Shields, K. A., Cook, K. D., & Greller, S. (2016). *How kindergarten entry assessments are used in public schools and how they correlate with spring assessments*. Washington, D.C.: Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/northeast/pdf/REL_2017182.pdf (2016).

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational researcher, 31*(7), 15-21.

Slavin, R. E., Cheung, A., Holmes, G., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal, 50,* 371-396. doi: 10.3102/0002831212466909

Snow, K. L. (2011). *Developing kindergarten readiness and other large-scale assessment systems: Necessary considerations in the assessment of young children*. Washington, DC: National Association for the Education of Young Children. (2011).

Supovitz, J. (2013). *The linking study: An experiment to strengthen teachers' engagement with data on teaching and learning*. Consortiium for Policy Research in Education.

Vitaro, F., Brendgen, M., Larose, S., & Trembaly, R. E. (2005). Kindergarten disruptive behaviors, protective factors, and educational achievement by early adulthood. *Journal of Educational Psychology, 97*(4), 617. ().

Wayman, J. C. & Jimerson, J. B. (2014). Teacher needs for data-related professional learning. *Studies in Educational Evaluation, 42*, 25-34.

Weisenfeld, G. G. (2017). Assessment tools used in kindergarten entry assessments (KEAs): State scan. New Brunswick, NJ: Center on Enhancing Early Learning Outcomes. Retrieved from http://ceelo.org/wp-content/uploads/2017/01/ceelo_fast_fact_kea_state_scan_2017_01_for_web.pdf

Table 1. Scalability of Best Practices

Consultation Feature	Supported Best Practices	Modifying for Scale	Potential Benefits	Trade-offs/ Considerations
One-on-One	• Individualization • Focused on discrete skills	Group delivery	• Fewer sessions needed • Opportunity for collaboration with peers • Changes to team processes/culture	• Inability to tailor to each teacher • Team dynamics • Less teacher engagement
In-Person	• Job-embedded	Web conference	• Easier to implement across large geographic areas • Greater flexibility in scheduling	• Less teacher engagement • Technology issues • Loss of personal connection
Ongoing	• Sustained • Timely	Stand-alone	• Fewer overall sessions needed • Greater flexibility in scheduling	• Not enough time to elicit change • Unable to apply skills in context between sessions

Table 2. School-Level Descriptive Statistics by Condition

	Control (n=73)	**Data Consultation** (n=59)
Student Characteristics		
% Below Benchmark		
Self-Regulation	16.61	18.10
	(7.34)	(9.48)
Social Skills	21.42	21.76
	(11.81)	(11.12)
Numeracy	12.29	12.13
	(7.95)	(7.29)
Patterning	28.41	24.92
	(14.25)	(12.71)
Geometry	26.56	24.17
	(14.25)	(12.93)
% Female Students	48.21	48.38
	(0.06)	(0.06)
% White Students	53.40	55.68
	(0.26)	(0.23)
% Attended Preschool	55.27	55.35
	(0.19)	(0.22)
% Disadvantage	50.34	50.81
	(0.18)	(0.18)
% Disability	7.16	7.96
	(0.18)	(0.05)
% English Language Learner	8.44	8.27
	(0.14)	(0.11)
Average Age (months)	65.03	65.10
	(0.60)	(0.66)
School Characteristics		
Accreditation Denied/Warned	0.23	0.17
	(0.43)	(0.38)
Average Class Size	18.67	18.66
	(2.95)	(2.46)
City	0.29	0.23
	(0.46)	(0.43)
Rural	0.45	0.52
	(0.50)	(0.50)

Note: Standard deviations reported in parentheses

Table 3. Covariate Balance[a]

	Data Consultation[b]	**In-Person[c]**	**One-on-One[d]**
School-Level Covariates[e]	(n=132)	(n=59)	(n=59)
Average Class Size	0.30	0.16	-0.23
	(0.23)	(0.32)	(0.36)
% Economically Disadvantaged	0.10	-0.09	0.44+
	(0.17)	(0.28)	(0.24)
% Special Education	-0.10	-0.31	0.07
	(0.18)	(0.27)	(0.26)
% White	-0.03	-0.50*	-0.39+
	(0.17)	(0.24)	(0.23)
Accreditation Status[f]	0.06	0.02	-0.16
	(0.09)	(0.14)	(0.13)
Rural	0.08	0.23	-0.31*
	(0.09)	(0.13)	(0.12)

+ $p<0.10$, * $p<0.05$

[a] Each covariate was regressed on an indicator of treatment status with fixed effects for randomization blocks. All results report the coefficient on the treatment indicator.

[b] First-stage randomization to data consultation or business-as-usual

[c] Second-stage randomization to in-person or web-based consultation delivery

[d] Second-stage randomization to one-on-one or group consultation format

[e] All continuous covariates reported in standard deviations

[f] Indicates whether or not the school was accredited in the previous school year

Robust standard errors clustered by school reported in parentheses.

Table 4. Differential Attrition at the Cluster and Teacher-Levels for Control vs. Any Treatment

	Intervention	Control	Overall
Cluster-Level Attrition			
Initial number of schools	59	75	134
Number of schools with survey responses	56	67	123
Cluster attrition rate	5%	11%	7%
Differential attrition	--	--	6%
Teacher-Level Attrition			
Number of teachers assigned	288	352	640
Number of teachers in clusters with responses	265	330	595
Number of teachers responding	131	189	320
Teacher attrition rate	51%	43%	46%
Differential attrition	--	--	8%

Table 5. Predictors of Treatment Take-Up

	Odds Ratio	**P-value**
% Below Benchmark - Self-Regulation	0.93	0.210
	(0.05)	
% Below Benchmark - Social Skills	1.09	0.104
	(0.06)	
% Below Benchmark - Numeracy	0.82	0.044
	(0.08)	
% Below Benchmark - Patterning	1.06	0.241
	(0.05)	
% Below Benchmark - Geometry	1.03	0.658
	(0.06)	
Did not complete EMAS	0.002	0.013
	(0.01)	
Not accredited	3.93	0.175
	(3.96)	
% White	3.57	0.598
	(8.61)	
% Attended Preschool	0.00	0.004
	(0)	
% Economically Disadvantaged	0.04	0.257
	(0.11)	
Student age	1.08	0.905
	(0.73)	
Class size	0.80	0.257
	(0.17)	
City	0.62	0.716
	(0.82)	
Rural	2.08	0.445
	(2.00)	

Note: Standard errors reported in parentheses.

Table 6. Results of Data Consultation vs. Business-As-Usual

	SDU - Confidence	**SDU - Collaboration**	**COA - Assessment Improves Education**	**Data Skills Assessment**	**Satisfaction with EMAS**	**Satisfaction with CBRS**
Data Consultation	0.19	0.21	0.16	-0.16	-0.03	-0.09
	(0.12)	(0.13)	(0.13)	(0.17)	(0.15)	(0.13)
Constant	0.81	0.79	0.18	0.99	0.98	0.85
R^2	0.04	0.05	0.05	0.15	0.08	0.07
N	273	273	277	217	281	285

+ $p<0.10$, * $p<0.05$

Note: All models included fixed effects for randomization blocks and covariate controls.

Robust standard errors clustered by school reported in parentheses.

All outcome variables are standardized.

Table 7. Results of Factorial

	SDU - Confidence	**SDU - Collaboration**	**COA - Assessment Improves Education**	**Data Skills Assessment**	**Satisfaction with EMAS**	**Satisfaction with CBRS**
One-on-One[a]	0.02	-0.14	0.14	0.10	0.01	-0.03
	(0.08)	(0.14)	(0.12)	(0.11)	(0.13)	(0.09)
In-Person[a]	-0.03	-0.02	0.06	-0.10	-0.20	**-0.19***
	(0.11)	(0.13)	(0.13)	(0.19)	(0.13)	(0.09)
Interaction[b]	0.07	-0.10	-0.03	**-0.24***	0.13	0.12
	(0.09)	(0.14)	(0.11)	(0.12)	(0.12)	(0.08)
Constant	0.92	0.81	0.13	0.18	0.46	0.25
R^2	0.23	0.14	0.11	0.31	0.25	0.22
N	105	105	106	90	107	111

* $p<0.05$

Note: All models included fixed effects for randomization blocks and covariate controls.

Robust standard errors clustered by school reported in parentheses.

All outcome variables are standardized.

Table 8. Exploratory Analysis - Business-As-Usual vs. Different Data Consultation Conditions

	SDU - Confidence	**SDU - Collaboration**	**COA - Assessment Improves Education**	**Data Skills Assessment**	**Satisfaction with EMAS**	**Satisfaction with CBRS**
In-Person, Group	0.05	0.18	0.15	-0.10	-0.23	-0.21
	(0.20)	(0.18)	(0.22)	(0.24)	(0.22)	(0.21)
In-Person, One-on-One	**0.36+**	0.03	**0.44***	-0.35	-0.08	-0.20
	(0.21)	(0.36)	(0.22)	(0.28)	(0.27)	(0.17)
Web-Based, Group	0.12	0.14	-0.03	-0.47	0.25	0.15
	(0.26)	(0.25)	(0.28)	(0.41)	(0.23)	(0.19)
Web-Based, One-on-One	**0.30+**	**0.38***	0.18	0.15	0.06	-0.06
	(0.16)	(0.17)	(0.16)	(0.18)	(0.24)	(0.22)
Constant	0.75	0.76	-0.21	1.01	0.93	0.84
R^2	0.04	0.05	0.06	0.18	0.09	0.08
N	273	273	277	217	281	285

+ $p<0.10$, * $p<0.05$

Note: All models included fixed effects for randomization blocks and covariate controls.

Robust standard errors clustered by school.

Figure 1. Data Use Cycle

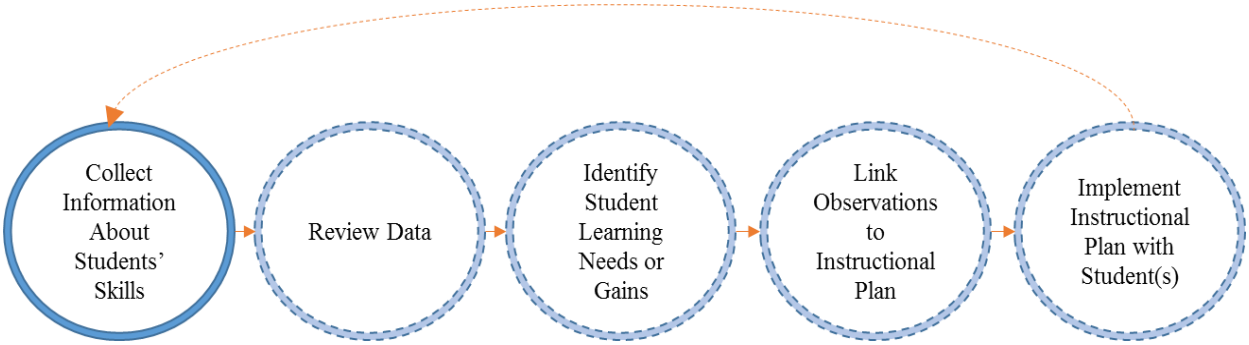
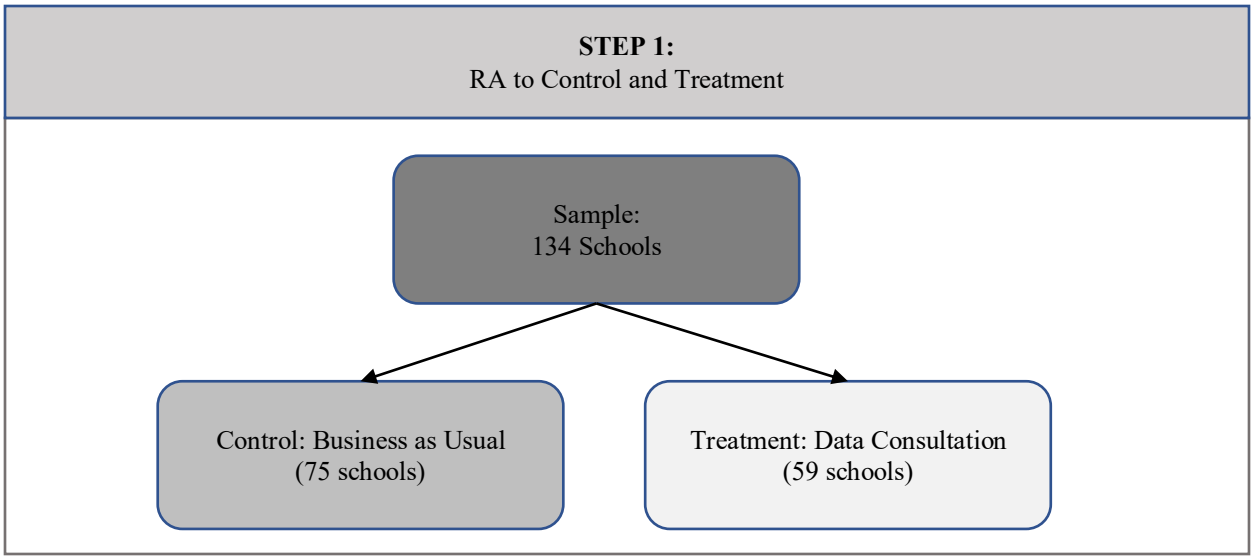
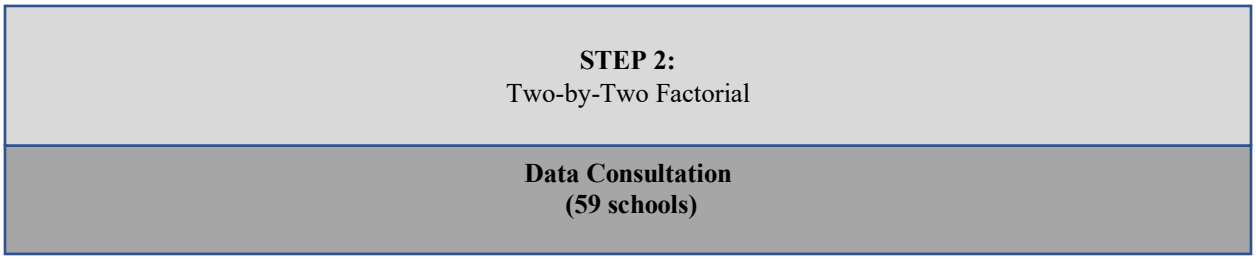


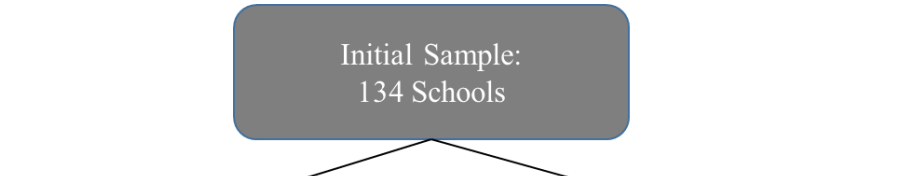
Figure 2. Two-Step Randomization Process





		Factor 1: Delivery		Factor 1: Delivery	
		Web-Based	Web-Based	In-Person	In-Person
Factor 2: Format	One-on-One	0.37	17 schools	-0.26	11 schools
	Group	-0.45	17 schools	-0.21	14 schools

Figure 3. Treatment Take-Up and Response Rates



			Factor 1: Delivery	
			Web-Based	In-Person
Factor 2: Format	One-on-One	Untreated:	10 (36%)	4 (40%)
		Treated:	7 (55%)	7 (57%)
	Group	Untreated:	8 (58%)	8 (45%)
		Treated:	9 (46%)	6 (41%)

The Experimental Switching Replication: A hybrid research design

Katherine L. Miller-Bains

Study 3: The Experimental Switching Replication: A hybrid research design

Abstract

Educational researchers face special challenges when making causal links between programs and their impacts. However, researchers can safeguard against common concerns through intentional research design. This paper discusses a hybrid research design, the experimental switching replication, which incorporates multiple design features including repeated measures, randomization, and replication in order to address threats to both internal and external validity. In an effort to provide practical guidance on how to implement the experimental switching replication, we discuss the required components of the design, highlight the strengths and weaknesses of the approach within an education context, and provide an example of how it was utilized within a teacher preparation program.

Introduction

Educational researchers are tasked with demonstrating how innovations improve teaching and learning. In order to make these connections, they conduct empirical studies to determine the effect of an intervention on its targeted outcomes. While the ideal evaluation would identify the "true" treatment effect across all relevant contexts, researchers face many challenges in establishing generalizable, causal relationships. Because researchers operate with constrained resources and large-scale experiments are very expensive, studies do not generally occur within the entire population of interest. Instead, researchers must balance the need to carry out evaluations in controlled, localized contexts in order to bolster internal validity, with the need to produce externally valid results that apply to all relevant samples and settings (Shadish, Cook, & Campbell,

2001). As the former frequently necessitates a narrow scope while the latter calls for a broader focus, the two forms of validity are frequently at odds with each other in a single study.

In the face of these tradeoffs, the education research community has prioritized studies that possess strong internal validity. This has led entities like the What Works Clearinghouse to deem experimental studies as the highest quality evidence of effectiveness and funding streams to incentive the use of highly controlled research designs. Conversely, there has been little emphasis on investigating how and if treatment effects replicate when administered using different samples, outcome measures, treatments, and settings in order to allow for externally valid inferences, with replications representing only 0.13% of studies published in top education journals (Makel & Plucker, 2014). This lack of replication has raised concerns as interventions that have performed well in small-scale, localized studies have failed to do so when introduced to more diverse samples and settings (Kraft, 2017).

Given the presumed tension between the two types of validity and constrained resources, it is not surprising that the prioritization of internally valid educational studies has coincided with a dearth of replication research. However, the internal and external validity of a study are not diametrically opposed nor does the inclusion of both require intensive increases in resources. Instead, different elements of research designs can be used in conjunction with one another to strengthen both types of validity within in a single study. Yet, such hybrid designs have been underexplored and underutilized in education research.

This paper explicates the potential benefits and limitations of one such research design – the experimental switching replication. Because this design incorporates the use of repeated outcome observations and randomization, it possesses strong internal validity, while the inclusion of additional treatment administrations improves external validity. We discuss the different components of the switching replication design using the causal replication framework and provide an example of an experimental switching replication within a teacher preparation program. This overview is intended to help researchers to prospectively plan experimental switching replications in order to conduct rigorous evaluations of educational programs.

Study Validity and Research Design

From the evaluation of educational interventions, researchers aim to draw inferences about the cause-and-effect relationship between a treatment and its targeted outcomes within a population of interest. Different studies, however, provide varying degrees of evidence to warrant these types of inferences. The extent to which a study supports claims about relationships between interventions and outcomes is referred to validity (Shadish et al., 2001).

Threats to Validity

Several factors often threaten the validity of educational evaluations. First, because schools and classrooms are dynamic contexts, concomitant events often influence teacher and student outcomes. The ability to rule out these alternative explanations for observed effects – referred to as *internal validity* (Campbell & Stanley, 1967; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2001) – is crucial for establishing the causal link from treatment to outcome. The stronger the internal validity

of a study, the more confident the researcher can be that intervention caused the observed effects. Secondly, researchers intend to make generalized claims concerning the relationships between treatments and outcomes, not just within the observed sample but across the larger population. This ability to apply inferences drawn from one study to other units, treatments, observations, and settings (Cronback, 1983) is referred to as *external validity*. Below, we describe specific threats to internal and external validity below.

Mortality/attrition. Over the course of a study, some of participants may drop out and fail to contribute outcome information. When characteristics of those participants missing information – such as motivation, health, interest, etc. – are also correlated with performance on the outcome, the posttest would reflect these differences absent any effect of treatment.

History. History threats occur when participants are exposed to events other than the treatment over the course of the study that result in changes to the target outcome. For instance, a school may experience several simultaneous, school-wide changes targeting students' math performance, such as teacher professional development, a new mathematics curriculum, and additional classroom time allocated to math instruction. Within this scenario, it is very difficult to isolate the unique contribution of one of these inputs to any observed improvements within the school.

Maturation. As educational interventions may take anywhere from several days to multiple years to implement, participants are likely to experience natural changes in their knowledge, skills, and abilities over the course of the evaluation. This development, referred to as maturation, encompasses "all of those biological or psychological processes

which systematically vary with the passage of time, independent of specific external events" (pp. 7-8, Campbell & Stanley, 1963). Researchers may erroneously attribute differences between earlier and later observations to treatment, when those changes would have occurred simply due to the passage of time, regardless of intervention delivery.

Testing Effects. When the act of completing a measure multiple times affects performance absent any intervention, the validity of the study can be threatened by testing effects. As test-takers become more familiar with the format and content of the measure, their scores can improve due to becoming more sensitized to errors or socially desirable responses, or decrease as a result of fatigue. Similarly, in the case of external observations, the presence of a rater or knowledge that they are being scored might influence a participant's performance. All of these changes in outcomes are the result of administering the test rather than the treatment.

Instrumentation. When measurement changes over the course of the study, observed effects may be attributable to differences in the instrument rather than treatment. As Shadish and colleagues clarify, instrumentation involves a change to the measure while testing reflects changes in the participant (pp. 60, 2001). Instrumentation is most frequently a concern when using administrative data or standardized tests, and changes are made to the measure (e.g., addition of new content, redefinition of constructs) that coincide with the intervention period.

Reactivity to Study. Participants may also respond to the fact they are in a research study in ways that have nothing to do with the treatment under evaluation. This threat is primarily a concern for those studies that incorporate a comparison group. In this

case, participants may be aware that they are not receiving something offered to others, and over-compensate (i.e., compensatory rivalry) or disengage (i.e., resentful demoralization) as a result. When this occurs, change over time reflects participants' responses to their role in the study and is no longer an accurate representation of the counterfactual.

Selection and the parallel trend assumption. Oftentimes, the manner in which participants receive treatment is outside of the researcher's control. Instead, factors such as motivation, skill, and personal interest determine who participates in the intervention and who does not. When the treated and untreated groups are non-equivalent in ways that are also related to performance on the outcome measure, the study is susceptible to the threat of selection. Selection may also interact with other threats to validity in ways that affect the trajectory of one group. Most commonly, these threats arise when comparing non-equivalent groups such that one group is exposed to different events that influence the outcome outside of treatment (i.e., selection-history interaction), improves at a faster or slower rate absent treatment (i.e., selection-maturation), or is more likely to experience floor or ceiling effects on a measure (i.e., selection-instrumentation).

Interactions between causal relationships with units, treatments, outcomes and settings. The aforementioned threats deal with the internal validity of study. However, if a researcher wishes to extrapolate the inferences made from one study to other relevant units, treatments, observation, or settings, then she must consider the likelihood that the established causal relationship is likely to interact with characteristics of the study sample. For instance, in many cases, schools elect to participate in a study, and it is from this smaller pool of that researchers conduct a randomized controlled trial.

However, it is likely that these schools differ from the broader population of interest in terms of need and their willingness to try a new innovation. This would limit the inferences that the researcher could draw from the small study to all relevant schools.

Using Research Design to Address Threats to Validity

Campbell and colleagues (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2001) have suggested the best way to safeguard against threats to validity is through the design of a research study. Because researchers are often constrained in some ways, whether by practical, political, or ethical considerations, certain approaches are often infeasible within the given parameters of a study. For instance, it may not be possible to increase the sample available for given study. However, researchers can work within the study's limits to identify research designs that counteract the most likely threats to validity and invest in the necessary design elements accordingly. Although it is possible to control for factors directly after the study is completed, using research design can often eliminate the need to measure all potential confounders. Given the scope of external factors external that are likely to affect the treatment outcome within educational settings, using research design often provides stronger and more convincing evidence of a program's effectiveness compared to statistical adjustments.

Borrowing Campbell's notation, we outline each study design and its features over time. Here, each line represents the sequence of observations (O) and treatments (X) a group receives, and each line represents a treatment group. We use an "R" to denote randomization of groups to treatments or treatment administrations.

Repeated measures designs. When the outcome of interest is collected is collected prior to and after the introduction of treatment, the researcher may use a repeated measures design to evaluate an intervention. Because of the flexibility and advantages of the repeated measures approach, these designs have been widely used to evaluate programs across the behavioral and educational sciences. In particular, special education research has long used the single-subject design as a means of improving upon the typical case study to evaluate and establish evidence-based practices (Horner et al., 2005; Richards, 2018). Within this context, repeated measures designs are especially appealing as it is often difficult to identify an appropriate comparison group for exceptional learners or withhold a potentially beneficial treatment from eligible participants. Education economists have also used variations of the repeated measures design, such as the interrupted time series and difference-in-differences, in order to evaluate changes in large-scale policies and programs (Dee, Jacob, & Schwartz, 2013; Grissom, Nicholson-Crotty, & Harrington, 2014). These types of studies have become more feasible as schools, districts, and states have improved their longitudinal data systems, allowing researchers to draw from administrative data to evaluate policy shifts (Hallberg, Williams, Swanlund, & Eno, 2018).

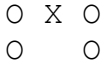
In its simplest form, the repeated measures requires only a pretest and a posttest collected within a single group.

O X O

The repeated measures approach is a natural fit for educational settings where assessments are commonly administered at regular intervals over the course of the school year, and it offers several advantages in contrast to cross-sectional designs, in which

outcomes are only observed at one point after treatment. First, because participants are measured both prior to and after intervention in a repeated measures study, the researcher is able to use pretreatment observations to estimate the counterfactual, or what would have occurred absent treatment. Treatment effects are then represented as the *change* within individuals. Since scores are generally more consistent within a person, the amount of outcome variance is reduced within a repeated measures framework. This means that the repeated measures designs offer improved statistical power and precision (Hedges & Hedberg, 2007), making it possible to detect treatment effects in smaller samples. While the simple pretest-posttest design provides an improvement over a cross-sectional comparison of posttest means and has been widely used in the social sciences, it is vulnerable to many threats to internal validity (Campbell & Stanley, 1963; Shadish, Cook, & Campbell, 2001). In particular, maturation, testing effects, history, and – in some cases – instrumentation offer plausible alternative explanations for observed treatment effects.

To better account for history threats, researchers commonly incorporate observations from an untreated comparison group. In this strain of repeated measures, the researcher identifies a group that would likely be exposed to many of the same historical events but did not receive the intervention. In educational studies, these groups often comprise students or teachers in the same or neighboring classrooms, schools, or districts that are likely to experience the same policies and programs outside of the intervention. The analog of the single-group pretest-posttest design is often referred to as a simple difference-in-differences or two-group pretest-posttest design.



Changes in the comparison group from the pre- to the post-treatment observation are intended to capture changes both groups would experience over time attributable to factors outside of treatment.

While the comparison group used within a difference-in-differences design need not be equivalent to the treated group, the two groups are assumed to follow parallel trends in order to produce valid estimates. In fact, the groups may be fundamentally dissimilar, but the differences between groups should be consistent across time prior to the treatment. For instance, a study of a remedial math program may use higher-achieving students who were ineligible for treatment as the comparison group. In this case, the two groups start at different levels of math achievement. However, it is assumed that students' math skills in both groups are improving at the same rate over time, absent the treatment. This assumption can be difficult to satisfy when individuals self-select into the intervention, as it can be difficult to know if baseline differences may interact with time-varying threats such as maturation or history.

Randomized designs. In order to ensure equivalence between groups, the researcher may randomize the treatment to participants. Randomized experiments have become more common in educational research in recent decades as they possess strong internal validity and causal rigor. Additionally, because selection mechanisms within educational settings are complex and difficult to quantify, randomized controlled trials are regarded as providing the most definitive evidence of a program's effectiveness as they eliminate the need to control for these factors directly.

In its simplest form, the randomized experiment assigns individuals to receive an intervention, and a posttest is collected after the administration of treatment.

```
R   O
R X O
```

This experimental designs provides strong causal warrant as the only difference between groups on expectation is the offer of treatment. Frequently, researchers will also collect a pretest prior to the administration of treatment.

```
R O   O
R O X O
```

This allows them to determine directly the equivalence between the treatment and control groups, and as discussed previously, the use of repeated outcome observations generally provides improved statistical power as the pretest explains variation in the outcome measure (Hedges & Hedberg, 2007).

While randomized experiments are considered the gold standard in educational evaluations, they present several challenges and considerations that can make them difficult to implement in educational settings. Foremost, these designs generally require withholding treatment from a subset of the sample. Because of this, they are best suited to situations where resources will not permit all eligible participants to receive treatment at the same time. Furthermore, because participants must relinquish control over who gets what and when, experiments are less politically feasible, so researchers often deal with limited samples in experimental studies. The also require the satisfaction of several assumptions in order to produce causally-interpretable results, primarily the assumption that treatment assignment does not influence participants potential outcomes and participants do not differentially drop out of the treatment conditions. Finally, because of the use of samples that are small and often unrepresentative of the entire population of interest, randomized experiments generally have low external validity.

Replication designs. Conceptually, replication focuses on determining whether the results from one study can be reasonably reproduced in other samples, settings, times, dosages, or research designs. In practice, it involves two or more trials of the same intervention conducted across these different variables. Previously, replication has largely been conceived of procedural or theoretical rather than a design element (Schmidt, 2009). However, Wong and Steiner (2018) have recently formalized the replication as a research design. Within their Causal Replication Framework, the researcher can either perform an exact replication or systematically vary one component of the original study in the replication, such as the sample, while holding all other elements constant. This allows the research to determine whether or not the results replicate despite or as a result of the variable across the treatment administrations. For example, a researcher may be interested in knowing if two different measures of the same outcome produce reasonably similar results. If so, they could conduct an experiment in a sample and provide the two measures to all participants. Because the respondents would have completed the same measures at the same time after receiving the same treatment, any differences could be attributed to differences in the instrument. This would then provide evidence that the intervention impacted the targeted construct rather than being an artifact of a single measure's properties.

In addition to improving external validity, the use of replication designs has the advantage of allowing all participants receive treatment over the course of the study, even within an experimental context. In one such design commonly referred to as the stepped wedge design or dynamic waitlist design (Brown & Lilford, 2006; Wyman, Henry,

Knolbauch, & Brown, 2015), groups are randomly assigned to the timing of a treatment over the duration of the study.

R O O O X_a O
R O O X_a O X_a O
R O X_a O X_a O X_a O

Because power is enhanced by increasing the number of treatment administrations or "steps" rather the number of participants or clusters, the stepped wedge is best utilized in situations that allow for a long study period is possible as well as many administrations of treatment (Fok, Henry, & Allen, 2015). Similarly, the stepped wedge is well-suited to studies of treatments that are assumed to have a lasting effect or that cannot be removed after it is introduced.

The Hybrid Research Design: The Experimental Switching Replication

The experimental switching replication combines the design elements described above by incorporating multiple observations, randomization, and multiple treatment administrations. Under this basic design, one group receives treatment earlier in the study while the other group serves as the control. The groups then switch roles in the second treatment administration, with the previously treated group serving as the control.

R O X_a O O
R O O X_a O

Much like the stepped wedge, participants are randomly assigned to groups that receive treatment at different points over the course of the study in the experimental switching replication design. However, the power no longer comes from the steps as treatment is removed after each trial, making each treatment administration independent. This also means the experimental switching replication can be executed over shorter period of time relative to the stepped wedge.

The use of all three design elements provides several advantages within the experimental switching replication over using repeated measures, randomization, or replication alone. First, the use of randomization mitigates any concerns that the two groups do not share a common trend over time, as random assignment ensures that the groups are equivalent on both observed and unobserved characteristics. Additionally, because treatment assignment is random and the trials are independent, the switching replication offers power advantages above and beyond what is provided by other repeated measures designs, as the two treatment administrations can be stacked and analyzed collectively (Edmonds & Kennedy, 2017). Including the replication component not only permits all participants to receive treatment, it allows the researcher to determine if the effects of treatment replicate in a second, equivalent sample at a different point in time.

While the switching replication has seen more widespread use in the harder sciences such as engineering, there are no examples in educational settings and few in the social sciences that we could identify. The few of the switching replication within these fields have been quasi-experimental (e.g., using repeated measures and replication but no randomization) (Bouwer, Koster, & van den Bergh, 2018; Stoddard & Piquette, 2010) and focused primarily on large scale policy evaluation (Parker, 1983; West, Hepworth, McCall, & Reich, 1989). In addition to the limited practical examples, few people have written about the mechanics design to date (Hedayat and Yang, 2005), barring a few pages in methodological texts and articles (Edmonds & Kennedy, 2017; Mercer et al., 2007; Shadish et al., 2001). Because of the lack of guidance, the experimental switching replication is not well understood.

Requirements of the design. While the experimental switching replication offers several advantages over more commonly used research designs, it also imposes several data requirements that may be difficult to satisfy in some educational contexts. Some requirements address the common validity threats outlined previously, while others stem from the additional assumption that each trial within the switching replication is independent. It is this underlying assumption allows the experimental switching replication to improve the evidence of external validity.

Attrition and study reactivity. Like other longitudinal studies, the switching replication has a greater risk of losing participants over the course of the study. Additionally, because the switching replication requires multiple treatment administrations during which some participants do not receive the intervention, differential attrition and study reactivity become greater concerns relative to traditional pretest-posttest randomized experiments. This means that it is best suited for situations with low risk of participant attrition over the course of the study. Classroom contexts often satisfy these requirements as students are able to complete multiple interventions and outcome assessments over the course of semester or academic year.

Carryover and sequence effects. As the trials are assumed to be independent of one another within a switching replication, the effects of the treatment in one period should not influence outcomes later in the trial. Because of this, the switching replication is best used to study treatments with immediate, short-term effects on the outcome. If the treatment effects are suspected to carryover into later trials, it becomes difficult to assume that the treatment groups are following the same trend later in the study. Similarly, when the experimental switching replication is used to evaluate more than one intervention

over the course of the study, it is assumed that the order in which participants received the treatments have no bearing on the estimated effectiveness. It is possible to account for some of these complications prospectively. For instance, if the researcher anticipates that the treatment will demonstrate a delayed or long-term effect, additional measurement periods should be added after each treatment administration in order to capture longer-term outcomes for both groups. Additionally, if sequence effects are likely, the researcher can counterbalance the ordering of the treatments such that each possible sequence is represented among the treatment groups (Ellis, 199; Kennedy & Edmonds, 2016).

Treatment contrast across trials. Much like the receipt of treatment in one trial should not influence the results of another trial within the experimental switching replication, the treatment contrasts should remain consistent across time. If the intervention changes substantially over time such that participants in later trials receive a different intensity or dosage, then it becomes difficult to know if the presence or absence of effects across administrations is due to differences in the treatment or a lack of replicable results.

Properties of outcome measure. As is the case with other repeated measures designs, the researcher must consider the properties of the outcome measure within the switching replication. First, because the treatment effect is estimated as a change within individual, the outcome must be able to be repeated and sensitive to detect differences over time. Second, highly variable outcomes are not well-suited for many repeated measures designs, as the design loses its efficiency. Additionally, while trends do not need to be linear, the researcher will need additional outcome observations in order to

properly detect and account for non-linear patterns. Lastly, the researcher must be able to observe the outcome at regular intervals over the course of the study.

Case Study: Switching Replication in Teacher Preparation

In order to demonstrate how the switching replication may be used within an educational context, we detail its application within a teacher preparation program below. The present multi-arm, experimental switching replication is part of a yearlong study to evaluate different skill-based coaching supports administered in a simulated learning environment in a school of education. In order to understand the study design as well as the rationale between the switching replication, we describe the simulator, the motivation for evaluating coaching supports within this context, and the intervention.

Overview of Simulator and Study

The simulator utilizes an online application to display avatars within a variety of education-related settings such as elementary and secondary classrooms. These classroom environments are projected onto a screen, and avatars serve as "virtual puppets" controlled by a trained specialist. This simulator specialist responds to participants' actions based on predefined character profiles and a detailed description of each scenario, allowing for standardization of responses across interactions with teacher candidates. While simulated learning environments have not been widely used in teacher education, they have proven to be a valuable tool to train professionals in other fields such as aviation (Hayes, Jacobs, Prince, & Salas, 1992) and medicine (McGaghie et al., 2006).

A unique and powerful feature of simulated learning is the ability for experts to directly observe novices' practices in a controlled environment, as opposed to less controlled field settings. This allows for participants to receive immediate, real-time

feedback on their performance as they complete a simulation. As such, the larger study is focused on contrasting different feedback conditions administered in the simulator and their relative effects on teacher candidates' targeted skills/behaviors when given an opportunity for repeated practice.

Sample and Context. Teacher candidates in the second year of a teacher preparation program used the simulator at three points over the course the academic year. As part of their general teaching methods course, the teacher candidates used the simulated learning environment to practice different skills covered in class. During the fall semester, teacher candidates focused on providing high-quality feedback to the virtual students during the discussion of an assigned text. In the spring semester, teacher candidates generated classroom expectations and norms with the avatars and were given opportunities to practice classroom management through redirection.

Intervention. In order to support the development of their practice in the simulator, we investigated different skills-based coaching formats. We randomly assigned teacher candidates to each of the following feedback conditions:

1. Self-reflection (treated as our control condition), in which the participants completed graphic organizers prompting them to evaluate what went well, what they would do differently, and which actions they planned to do in the second iteration;
2. Guided reflection with a coach, in which the participants received directive feedback on their performance at the end of the simulation;
3. Guided reflection plus real-time coaching, in which a coach used quick prompts delivered via an earpiece to highlight opportunities for the participant to use the

targeted skill as they occurred in during the simulation in addition to directive feedback at the end of the simulation. For example, coaches would say "voices off" when an avatar began singing aloud, indicating that the teacher should redirect the virtual students' behavior.

All sessions were video recorded, and trained research assistants coded videos and/or transcripts of teacher candidates in the simulator.

Rationale for Switching Replication

While we wished to evaluate the effects of coaching in the simulator in order to inform future provision of such supports, executing an experimental or quasi-experimental study in the teacher preparation program posed practical and ethical concerns. Both teacher candidates and program faculty expressed reluctance about withholding treatments from a subset of study participants. Yet, we had little empirical evidence to suggest that skills-based coaching improved teacher candidates' skills, making it difficult to justify the intensive investments needed to provide coaching to all teacher candidates each time they used the simulator. Because the students were exposed to many other over the course of the academic year, it would be difficult to rule out history threats without the use of a comparison group. Given concerns about withholding treatment and opportunities for more than one exposure to treatment and subsequent measurement, we implemented a switching replication to both counter limitations of alternative designs and capitalize on the inherent characteristics of our study sample.

Study Design

Because of our interest in multiple treatments, timeframe, and some logistical challenges, we modified the traditional switching replication design presented in Table 1. Instead, our study took the following form:

R		O_1	O_2	X_g	O_3
R	X_g	O_1	O_2	X_r	O_3
R	X_r	O_1	O_2		O_3

Here, we randomized participants to one of the three groups blocking by the four fall course sections (two elementary and two secondary sections). In the fall treatment administration, the first group received self-reflection (no-treatment control group), the second group received guided reflection (X_g), and the third group received real-time coaching in addition to guided reflection (X_r). After receiving treatment, we observed the teacher candidates performance in the simulator (O_1). Following a brief washout period, we collected a midpoint observation (O_2) prior to implementing the second treatment. This observation serves as second baseline as no treatment was provided between O_1 and O_2. During the second trial, all participants switched to a different treatment, with those who had been assigned previously to self-reflection receiving guided reflection, those who had received guided reflection receiving real-time coaching, and those who had received real-time coaching switching to self-reflection in the spring. After receiving these treatments, we again observed their performance in the simulator (O_3).

Complications and Considerations

In designing our study, we hoped to provide strong empirical evidence regarding the effectiveness of coaching supports within a limited sample. However, we must also consider the implications of the current context on the validity and interpretability of our results. While the experimental switching replication safeguards against many threats to

internal validity such as history and the presence of parallel trends, it is still vulnerable to differential attrition and study reactivity. We anticipated that attrition would be less problematic over the course of the study as teacher candidates were expected to use the simulator as part of their academic program. Additionally, we did not anticipate any residual effects from the first trial, as we targeted different skills in the fall and spring semesters and included a washout period of 2 months before the second treatment administration. Finally, the basic parameters of the intervention (e.g., self-reflection, coaching, and coaching plus real-time feedback) remained the same across administrations over the course of the study, though several aspects of the coaching intervention changed such as the targeted construct. However, the duration of these interactions and the number of suggestions remained the same.

Discussion

Educational evaluators are expected to make convincing connections between programs and outcomes. As a result, the research community has come to prioritize studies with strong internal validity, and the traditional randomized experiment has been held in the highest regard due to its ability to rule out alternative explanations for observed treatment effects. However, traditional applications of the randomized experiments generally lack the ability to produce externally valid results. Alternatively, non-experimental cross-sectional studies frequently struggle to rule out competing explanations for observed effects, making it difficult to link treatments to outcomes. Given the rising standard for the research produced in educational contexts and the difficulty of executing traditional randomized controlled trials, it is imperative that researchers have more methodological tools that can multi-task – providing results with

strong causal warrant while also providing confidence that the treatment effects are more than a byproduct of chance or specific conditions.

Though the repeated measures design has been used to evaluate an array of educational program and policies, the switching replication variant is not well represented both in the methodological and applied literature. Without practical guidance, this design is likely to remain underutilized in educational evaluations. However, this design offers many of the strengths of the randomized experiment while also providing additional advantages such as improved external validity and the ability to provide treatment to all study participants. It also conducive to many educational settings where resources constrain the implementation of new supports to all participants simultaneously, but withholding treatments would not be desirable. The classroom often provides an ideal context for using the switching the replication, as it limits the risk of differential attrition and provide opportunities for repeated treatment administrations and outcome observations. Furthermore, as longitudinal data become more common across educational settings, the more feasible these types of studies become.

Future Directions

It is important for the researcher to weigh the most likely threats within the study and make investments accordingly. As we discovered through our application of the experimental switching replication design within a teacher preparation program, it can be difficult to satisfy the primary assumption of independence across treatment administrations. Still, while these divergences for the intended design complicate the interpretation of the results, some of these issues can be addressed in the analysis of the data. Future work summarize the different analytic approaches that can be employed to

estimate treatment effects from the switching replication, further sensitivity checks that can be used to test the underlying assumptions.

References

Barry Issenberg, S., McGaghie, W. C., Petrusa, E. R., Lee Gordon, D., & Scalese, R. J. (2005). Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. Medical teacher, 27, 10-28. doi: 10.1080/01421590500046924

Basadur, M., Graen, G. B., & Scandura, T. A. (1986). Training effects on attitudes toward divergent thinking among manufacturing engineers. *Journal of Applied Psychology*, *71*(4), 612.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reform. Evaluation Review, 27, 3-49. doi: 10.1177/0193841X02239017

Bouwer, R., Koster, M., & van den Bergh, H. (2018). Effects of a strategy-focused instructional program on the writing quality of upper elementary students in the Netherlands. *Journal of Educational Psychology*, *110*(1), 58.

Brown, C. A., & Lilford, R. J. (2006). The stepped wedge trial design: a systematic review. *BMC Medical Research Methodology*, *6*(1), 54.

Campbell, D. T. & Stanley, J. C. (1963). Experimental and quasi-experimental design for research. Handbook for Research on Teaching. Boston, MA: Houghton Mifflin Company.

Cook, T. D. & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston, MA: Houghton Mifflin.

Cernin, P. A., & Lichtenberg, P. A. (2009). Behavioral treatment for depressed mood: A pleasant events intervention for seniors residing in assisted living. *Clinical Gerontologist*, *32*(3), 324-331.

Dee, T. S., Jacob, B., & Schwartz, N. L. (2013). The effects of NCLB on school resources and practices. *Educational Evaluation and Policy Analysis*, *35*, 252-279. doi: 10.3102/0162373712467080

Edmonds, W. A., & Kennedy, T. D. (2017). *An applied guide to research designs: Quantitative, qualitative, and mixed methods.* Los Angeles, CA: SAGE Publications.

Ellis, M. V. (1999). Repeated measures designs. The Counseling Psychologist, 27, 552-578. doi: 10.1177/0011000099274004

Fanning, R. M., & Gaba, D. M. (2007). The role of debriefing in simulation-based learning. *Simulation in healthcare*, *2*, 115-125. doi: 10.1097/SIH.0b013e3180315539

Feng, J. Y., Chang, Y. T., Chang, H. Y., Erdley, W. S., Lin, C. H., & Chang, Y. J. (2013). Systematic Review of Effectiveness of Situated E-Learning on Medical and Nursing Education. *Worldviews on Evidence-Based Nursing*, *10*(3), 174-183.

Fok, C. C. T., Henry, D., & Allen, J. (2015). Research designs for intervention research with small samples II: Stepped wedge and interrupted time-series designs. *Prevention Science*, *16*(7), 967-977.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job

attitudes. *Educational Evaluation and Policy Analysis*, *36*(4), 417-436. doi: 10.3102/0162373714533817

Hallberg, K., Williams, R., Swanlund, A., & Eno, J. (2018). Short comparative interrupted time series using aggregate school-level data in education research. Educational Researcher, 47, 295-306. doi: 10.3102/0013189X18769302

Hedayat, A. S., & Yang, M. (2005). Optimal and efficient crossover designs for comparing test treatments with a control treatment. *The Annals of Statistics*, *33*(2), 915-943.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional children*, *71*, 165-179. doi: 10.1177/001440290507100203

Marcantonio, R. J. & Cook, T. D. (1994). Convincing quasi-experiments: The interrupted time series and regression-discontinuity designs. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of Practical Program Evaluation*. San Francisco: Jossey-Bass.

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2006). Effect of practice on standardised learning outcomes in simulation-based medical education. *Medical Education*, *40*, 792-797. doi: 10.1111/j.1365-2929.2006.02528.x

McGaghie, W. C., Issenberg, S. B., Petrusa, E. R., & Scalese, R. J. (2010). A critical review of simulation-based medical education research: 2003–2009. *Medical Education*, *44*(1), 50-63.

Mercer, S. L., DeVinney, B. J., Fine, L. J., Green, L. W., & Dougherty, D. (2007). Study designs for effectiveness and translation research: identifying trade-offs. *American journal of preventive medicine*, *33*(2), 139-154.

Metzler, C. W., Biglan, A., Rusby, J. C., & Sprague, J. R. (2001). Evaluation of a comprehensive behavior management program to improve school-wide positive behavior support. *Education and treatment of Children*, 448-479. doi: https://www.jstor.org/stable/42900503

Palmer, S. B., Wehmeyer, M. L., Gipson, K., & Agran, M. (2004). Promoting access to the general curriculum by teaching self-determination skills. *Exceptional Children*, *70*(4), 427-439.

Parker, E. B. (1963). The effects of television on public library circulation. *Public Opinion Quarterly*, *27*(4), 578-589.

Richards, S. B. (2018). *Single Subject Research: Applications in Educational Settings*. Belmont, CA: Cengage Learning.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90.

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

Stoddard, H. A., & Piquette, C. A. (2010). A controlled study of improvements in student exam performance with the use of an audience response system during medical school lectures. *Academic Medicine, 85*(10), S37-S40.

West, S. G., Hepworth, J. T., McCall, M., & Reich, J. W. (1989). An evaluation of Arizona's July 1982 drunk driving law: Effects on the city of Phoenix. *Journal of Applied Social Psychology, 19*(14), 1212-1237.

Wong, V. C., & Steiner, P. M. (2018). Designs of empirical evaluations of nonexperimental methods in field settings. *Evaluation Review*, doi: 0193841X18778918.

Wyman, P. A., Henry, D., Knoblauch, S., & Brown, C. H. (2015). Designs for testing group-based interventions with limited numbers of social units: The dynamic wait-listed and regression point displacement designs. *Prevention Science*, *16*(7), 956-966.

Table 1. Overview of Switching Replication Design

1. Basic Design: Repeated measures design in which one group receives treatment while the other group serves as the comparison. These groups switch roles at a later point during the study.
2. Design Components
    - Treatment is withheld from each group at one point in the study.
    - The researcher collects repeated measures of the outcome variable before and after each treatment administration.
    - The treatment is administered at more than one point during the study.
3. Benefits
    - All participants receive the treatment
    - Improved statistical precision through repeated measures
    - Ability to assess replicability of results across administrations
4. Considerations of Switching Replication
    - Does the current context allow for multiple administrations of measures and treatment?
    - What is the risk of attrition over the course of the study?
    - Are there carryover effects after treatment is removed?
    - Are the treatment effects likely to be delayed?
5. Variants of the Switching Replication
    - *Experimental Switching Replication*: Randomly assigns individuals to two treatment groups where one receives the intervention first and the other receives treatment at a later time
        - Example: Cernin & Lichtenberg, 2009
    - *Multi-Arm Switching Replication*: Compares multiple treatment conditions
        - Example: Winterer,…, & Streffer, 2013
    - *Quasi-Experimental Switching Replication:* Does not use randomization to assign individuals to treatment groups, and instead uses intact groups such as classrooms, schools, states, etc.
        - Example: Basadur, Graen, & Scandura, 1986