HOW DO PRE-SERVICE TEACHERS LEARN? USING RIGOROUS RESEARCH METHODS TO INFORM TEACHER PREPARATION POLICY

A Dissertation

Presented to

The Faculty of the School of Education and Human Development

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Anandita Krishnamachari, M.Sc.

May 25, 2021

HOW DO PRE-SERVICE TEACHERS LEARN? USING RIGOROUS RESEARCH METHODS TO INFORM TEACHER PREPARATION POLICY

Anandita Krishnamachari, M.Sc.

Advisor: Vivian C. Wong, Ph.D.

EXECUTIVE SUMMARY

Teacher quality is the most important educational factor predicting student achievement (Darling-Hammond, 2000), and a large proportion of training effective teachers occurs during the preparation phase (Goldhaber, 2019). It is widely acknowledged that teacher preparation programs are tasked with providing pre-service teacher candidates with the skills and knowledge they need to effectively teach in future classroom contexts, although the extent to which this happens is debated (Greenberg & Jacobs, 2009). The literature suggests that novice teachers consistently report feeling underprepared and overwhelmed when they first begin to teach in a classroom environment (Goldhaber, 2018), and largely learn "on the job".

To improve the current state of teacher preparation, there needs to be a greater understanding of how preparation programs affect pre-service teachers' knowledge, skills and beliefs about teaching requires researchers and teacher educators to incorporate rigorous data collection and analyses into program structures. Although there is more quantitative research on teacher preparation in the last two decades than previously, these studies are plagued by "the quality of research design in the majority of these studies, including problems with small sample size, lack of control or comparison group, and

subjective outcome measures" (Mitchel & King, 2016). Overall, there is a need for more quantitative research that uses larger sample sizes, captures variation in program elements and teacher candidates, and more longitudinal research and multi-site studies that can causally link program components to teacher effectiveness. Measuring and using candidates' skills and knowledge during the preparation phase can bolster continuous improvement efforts at the program level and provide researchers with a better understanding of how prospective teachers learn during the preparation phase.

This dissertation seeks to address the challenges highlighted above in the current landscape of teacher preparation using rigorous quantitative research methods. The findings across chapters presented in this dissertation have important implications for how teacher preparation programs are designed, implemented and evaluated and contribute much-needed, rigorous quantitative evidence on how pre-service teachers learn to teach during the preparation phase. Together, the three chapters provide much needed evidence on the role of personal beliefs, attitudes and perceptions in how pre-service teachers learn during the preparation phase, the efficacy of providing them with targeted supports while they are learning and the promise of using robust data and rigorous research designs to inform teacher preparation policy.

Keywords: Teacher preparation, quantitative methods, education policy

Research, Statistics and Evaluation
School of Education and Human Development
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, (How Do Pre-Service Teachers Learn? Using Rigorous Research Methods To Inform Teacher Preparation Policy), has been approved by the Graduate Faculty of the School of Education and Human Development in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Name of Chair (Vivian C. Wong)

_____
Committee Member Name (Julie Cohen)

_____
Committee Member Name (Jim Soland)

_____
Committee Member Name (Jason Downer)

_____
Committee Member Name (Nathan Jones)

_____Date

DEDICATION

I dedicate this dissertation to my grandfather, my Boy, who always emphasized the importance of education, and would have been delighted to see me complete this PhD.

ACKNOWLEDGEMENTS

Ella Pugazhum Iraivanukke

This dissertation was made possible through the support and guidance of so many through the years. I would like to thank my committee, Dr. Jim Soland, Dr. Jason Downer and Dr. Nate Jones for their encouragement and thoughtful feedback that has only improved this dissertation tremendously. Your insights, critiques and questions have been extremely valuable in pushing this work along, and I look forward to many more discussions as I send these chapters out for peer review.

Thank you to all the members of the TeachSIM, SERA and Replication Collaboratory Labs. To each interactor, coach, research assistant and fellow graduate student, thank you for your contributions! The promises of this dissertation would be incomplete without your role in the work we do. A big thank you to Christina Taylor, who has been a source of inspiration, encouragement and the most wonderful friend/colleague for the last few years. To Alexis Prijoles and Rose Sebastian, who I have learned so much from and have truly enjoyed working with, thank you.

I would also like to thank the members of EdPolicyWorks for their encouragement and support during my PhD, and for providing a collaborative community that I am and will always be proud to be a part of. Thank you for always being willing to listen, to offer advice and feedback and for being a family through all the highs and lows that come with a PhD. A special thank you to Leslie Booren and Shannon Kontalonis, who have been incredible at managing logistics and making sure I don't drop the ball on international student "to-dos" . Thank you also to my master's advisor, Caroline Wong, for your constant support and advice- you have pushed me to become a better person and scholar, and for that I am extremely grateful. Thanks also to Dr. Ezekiel Dixon-Roman, who told me that working with Dr. Vivian Wong would be a wise decision- I couldn't agree more!

A big thank you to all of my EPW friends and colleagues who have been supportive along the way. In particular, I would like to thank my cohort-mates, Justin Doromal, Kylie Anglin, Emily Wiseman and Joe Kush. I can't believe we made it to the end, and I am not sure I would have made it through without the goats. Your support, intelligence, wit and thoughtfulness have made this journey truly special. I'd also like to thank Dr. Jessaylnn James, Dr. Katharine Meyer, Dr. Preston Magouirk, Dr. Rebekah Berlin, Dr. Kate Miller-Bains and Dr. David Martin for their support, feedback and encouragement over the years. To the cohort members that follow- Brian Kim, Walter Herring, Arielle Boguslav, Eli Talbert and Todd Hall, thank you for making the student

office such a fun work space to be in, and for bringing wisdom, joy, knowledge and new perspectives with you. I am so excited to see what the future brings for each of you!

A special thank you to Dr. Aliza Husain who made me feel like I belonged in more ways than I can count- thank you for the walks, for the advice, for the support, and for just being there when I needed you. I will always be grateful. To Dr. Tutrang Nguyen, who has become more of a friend than a colleague, thank you for being so amazing. Every walk, every FaceTime call, every text conversation has made me appreciate you so much! This would not have been possible with you.

Finally, and most importantly, I'd like to thank my advisor, Dr. Vivian Wong and co-advisor, Dr. Julie Cohen. Words fail to describe how much you both mean to me- thank you for taking a chance on me! Your continued support, encouragement and mentorship have truly meant the world to me, and I am privileged to call you family. Thank you for always understanding, for telling me I could do it, and for pushing me to become a better person and scholar. Thank you for encouraging me to work lesser at times, think more, be more precise and translate my thoughts into writing in a concise, cohesive way. You have given me so much time, love and support, and I am so grateful! Thank you.

To my incredible family, who has stood by me from the beginning, thank you. To my Appa, who has always believed in me (more than I ever have), trusted me to make the right decisions, and pushed me to "do what you love, rather than being forced to love what you do". My Amma, whose wings I fly on- thank you for being my North star, my guide, best friend and confidante. I am so lucky to have both of you! To my brother, for pushing me to get a PhD, and for believing me in me so much more than I ever have. I love you, and this one is for you. Aarti, thank you for always supporting me, for showing me how it's done and for all the love. To my in-laws- Mom, Dad, Akka, Tai, Prabhu and Adira, thank you for supporting and encouraging me, for believing in me and celebrating all my accomplishments, however small. To my Periamma and Periappa, for being a second set of parents, and for the constant support and "good shows". You both mean the world to me!

Finally, to Pratap- for switching between the husband and advisor caps as I needed, for inspiring me to work harder and with more dedication, for always believing in me, and for the countless hugs and ice cream. This dissertation belongs to you as much as it does to me-thank you for being such an amazing partner.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER IV:

APPENDICES:

LIST OF FIGURES

APPENDICES:

CHAPTER I: INTRODUCTION

Teacher quality is the most important educational factor predicting student achievement (Darling-Hammond, 2000). Research examining the variation in achievement growth among students assigned to different teachers concludes that students consecutively assigned to highly effective teachers perform significantly better than students assigned to a series of ineffective teachers (Rockoff, 2004; Rivkin, Hanushek & Kain, 2005; Kane, Rockoff & Staiger, 2008). Although there is still debate on exactly what constitutes high-quality teaching, the field of teaching is coalescing around the idea that high-quality teaching depends on a "flexible repertoire of high-leverage strategies and techniques that can be deployed with good judgement depending on the specific situation and context" (Loewenberg Ball & Forzani, 2009; pg.503). In essence, teaching is a complex and multifaceted process that involves a combination of knowledge, skills, beliefs and personal factors.

A majority of the formal training on how to be an effective teacher occurs during the preparation phase, before teachers even step into their classrooms (Goldhaber, 2018). The central assumption supporting teacher preparation is that by offering teacher candidates a package of coursework and practice opportunities, programs provide candidates with flexible knowledge and skills that they can apply in future teaching contexts (Leinhardt, Young and Merriman, 1995). Along with providing teacher candidates with practical skills through coursework and methods-based training (see Grossman, Wilson & Shulman, 1989; Wilson, Shulman & Richert, 1987), teacher

preparation programs are also responsible for developing the more affective elements of teaching, including teachers' resilience, efficacy and dispositions that determine the kind of the teacher they will be (Shoffner, 2008; Stahl, Sharplin & Kehrwald, 2016). In a classroom context, the affective elements of teaching are as crucial to effective teaching as foundational pedagogical skills since teachers must be able to consistently connect content knowledge, pedagogy and curriculum with their own emotional understanding, beliefs about teaching and personal histories (Zembylas, 2007).

Teacher preparation and teacher quality

Recently, teacher preparation programs have been under scrutiny for failing to adequately prepare graduates to teach in classrooms. Criticisms include low standards for admission, weak content-focused preparation, and coursework focused too heavily on theory to be practically applicable (Greenberg & Jacobs, 2009). More concerningly, evidence suggests that most novice teachers learn "on the job" and report feeling underprepared by their preparation programs (Atteberry, Loeb & Wyckoff, 2015; Kraft & Papay, 2014). Data from surveys completed by novice teachers show that the majority feel ill-prepared to teach, suggesting that there is potential for improvement among teacher preparation programs (Goldhaber, 2018).

Although teacher preparation programs are crucial to training high-quality teachers, researchers struggle to causally link elements of teacher preparation to teacher practice (Levine, 2006; Wilson, Floden & Ferrini-Mundy, 2001). Quantitatively, the most common approach to estimating the causal impact of teacher preparation programs on teacher quality is to approximate in-service teachers' "value-added" to student achievement once they actually enter the classroom (see Braun, 2015; Darling-

Hammond, 2015; Goldhaber, 2015) and tying these estimates back to training programs. While such estimates are helpful in understanding how the quality of preparation ultimately affects student outcomes, they do not provide actionable evidence that can be used to inform program design (Mitchel & King, 2016).

Understanding how preparation programs affect pre-service teachers' knowledge, skills and beliefs about teaching requires researchers and teacher educators to incorporate rigorous data collection and analyses into program structures. Although there is more quantitative research on teacher preparation in the last two decades than previously, these studies are plagued by "the quality of research design in the majority of these studies, including problems with small sample size, lack of control or comparison group, and subjective outcome measures" (Mitchel & King, 2016). Overall, there is a need for more quantitative research that uses larger sample sizes, captures variation in program elements and teacher candidates, and more longitudinal research and multi-site studies that can causally link program components to teacher effectiveness. Measuring and using candidates' skills and knowledge during the preparation phase can bolster continuous improvement efforts at the program level and provide researchers with a better understanding of how prospective teachers learn during the preparation phase.

How do pre-service teachers learn to teach?

One of the biggest gaps in the current literature on pre-service teacher education is evidence on how pre-service teachers learn to teach, and what factors influence this learning during the preparation period. This information is critical not only to understanding the impact that preparation experiences have on teacher candidates' instructional skills as well as beliefs, efficacy and attitudes towards teaching, but also

facilitates redesigning and strengthening existing teacher preparation programs to ensure that teacher candidates are ready from their first day in the classroom. Bandura's Social Cognitive Theory suggests that teacher candidates' learning trajectories are shaped by the interaction of three forces: their personal background (academic record, previous classroom experience, beliefs and attitudes towards teaching and personality characteristics among other variables), their formal training (components of the teacher preparation program they are enrolled in), and their behaviors (including teaching performance, self-regulation skills, etc.). These three factors—personal background, preparation, and behaviors—define teacher candidates' paths into the teaching profession. (Alt, 2015; Bandura, 1997a; Humphrey & Wechsler, 2006).

Each of these three factors and their interactions are depicted in Figure 1. At entry to the preparation program, teacher candidates bring with them a set of personal factors which collectively act as a filter through which pre-service teachers experience the teacher preparation and shape what they learn (Fang 1996; Lortie, 1975). These personal factors include attitudes and beliefs about teaching developed as a result of candidates' own experiences as students (termed by Lortie (1975) as the apprenticeship of observation), personality characteristics (such as neuroticism, extraversion, openness, etc.) and demographic characteristics (including gender, race, ethnicity, etc.). Once candidates begin coursework and accumulate teaching experiences during the preparation phase, their beliefs and attitudes are likely to evolve and change depending on how they cognitively process their development. At the same time, these beliefs and attitudes also act as filters through which pre-service teachers experience the preparation program (Fang, 1996; Ní Chróinín & O'Sullivan, 2014). Ultimately, teacher candidates' teaching

skills (or behaviors) are influenced by and in turn influence their personal factors, components of teacher preparation programs, and the interaction between them.

What is less clear from existing research on teacher preparation programs are the best ways to effectively prepare teacher candidates. There is evidence to suggest that the existing "one-size-fits-all" approach may not be useful in preparing effective teachers (Beijaard, Korthagen & Verloop, 2007; Howard & Milner, 2014) given that novice teachers report feeling ill-prepared for the classroom from their training experiences. Identifying and supporting teacher candidates in ways that improve their skills and beliefs about teaching during the preparation phase is critical, since training opportunities are limited once pre-service teachers enter the classroom (Palomera, Fernandez-Berrocal & Brackett, 2008). In addition, we know that novice teachers are at higher risk for experiencing stress and burnout (Hopkins, Hoffman & Moss, 1997; Roness, 2011) which leads them to leave the profession at disproportionately higher rates than their experienced peers (Roness, 2011).

Scope of this dissertation

This dissertation seeks to address the challenges highlighted above in the current landscape of teacher preparation using rigorous quantitative research methods. In the Chapter 2, I use hierarchical linear models to examine the variation in the development of self-efficacy beliefs among pre-service teachers enrolled in a university-based teacher preparation. Specifically, I focus on the role of teaching specialty, personality characteristics and attitudes towards teaching in determining self-efficacy beliefs. The goal of this chapter is to add critical (but missing) longitudinal evidence on the development of self-efficacy beliefs among pre-service teacher candidates during the

preparation phase. In Chapter 3, I use data from multiple experimental replication studies to examine the best ways to provide pre-service teacher candidates with standardized, simulated practice opportunities and targeted feedback (in the form of directive coaching) to improve their instructional skills. In particular, I assess whether the effects of providing coaching to improve pre-service teachers' pedagogical skills in a simulated classroom environment are robust across different sources of variation (including cohorts of pre-service teacher candidates, mode of delivery of coaching, different populations of interest and pedagogical task). Here, the goal is to provide teacher educators, preparation programs and policymakers with causal evidence on the contexts and conditions under which standardized practice opportunities and targeted coaching can be used to improve performance during the preparation phase. In Chapter 4, I present results from a randomized control trial evaluating the use of mental rehearsal techniques on pre-service teachers' stress levels and performance in a simulated teaching task focusing on classroom management. Through this chapter, I add to the limited research on how to best prepare teacher candidates to deal with the uncertainty and stress that accompanies teaching.

The findings across chapters presented in this dissertation have important implications for how teacher preparation programs are designed, implemented and evaluated and contribute much-needed, rigorous quantitative evidence on how pre-service teachers learn to teach during the preparation phase. Results from the second chapter highlight variation in how pre-service teachers perceive their own ability to teach effectively based on personal factors, including teaching specialty, personality characteristics and attitudes towards teaching. Acknowledging and examining this

variation and role of personal factors can provide teacher educators and preparation programs with information that can help design targeted supports and experiences to improve candidates' beliefs in their ability to teach as well as their teaching practices. Chapters 3 and 4 provide evidence on the use of targeted supports to improve teaching practices and beliefs among pre-service teachers and reduce feelings of stress and burnout. Chapter 3 provides evidence on the efficacy of standardized practice opportunities and coaching supports to improve pre-service teachers' practices during a teaching opportunity across different sources of variation. In addition to demonstrating the significant and positive impact coaching has on improving teaching skills, this chapter also shows that coaching is consistently effective across similar groups of teacher candidates who have the theoretical knowledge to support their practice, irrespective of the pedagogical task that they are practicing or the mode of delivery of coaching supports. Findings from the final chapter provide important evidence on the use of cognitive appraisal strategies such as mental rehearsal in facilitating the improvement of performance and perceptions among pre-service teacher candidates, as well as whether providing teacher candidates with a combination of coaching (problem-focused coping strategy) and mental rehearsal (emotion-focused coping strategy) helps teacher candidates cognitively re-appraise a stressful situation such as classroom management and improve their skills at managing a classroom environment and redirection off-task student behaviors. Together, these chapters provide much needed evidence on the role of personal beliefs, attitudes and perceptions in how pre-service teachers learn during the preparation phase, the efficacy of providing them with targeted supports while they are

learning and the promise of using robust data and rigorous research designs to inform teacher preparation policy.

Chapter 2: Examining the role of personal factors in the development of self-efficacy beliefs among pre-service teachers

At present, we know very little about how a teacher candidate's prior beliefs and attitudes or their personal characteristics influence their development during the teacher preparation phase. The literature suggests that all pre-service teachers bring their beliefs and perceptions regarding teaching and learning shaped through their own experiences as students (known as "the apprenticeship of observation") to their preparation program (Lortie,1975). There is also evidence that pre-service teachers generally perceive learning through the lens of their prior knowledge, preconceptions and beliefs (Ethell & McMeniman, 2000), and that these beliefs act as filters through which pre-service teachers experience the preparation program (Fang, 1996; Kagan, 1992; Ní Chróinín & O'Sullivan, 2014). Acknowledging and recognizing the role of personal factors in how pre-service teacher candidates learn to teach can provide both preparation programs and researchers with actionable information on how to improve candidates' preparation experiences to prepare them for the classroom.

Yet, the research on how prior perceptions and beliefs influence the development of teaching skills has not extended beyond when a teacher candidate first enters the preparation program. Most of the existing research on pre-service teacher candidates' characteristics has focused on admission criteria including undergraduate GPA, SAT/ACT score and selectivity of undergraduate institution attended (Mitchel & King, 2016) and linking these criteria to student outcomes to get a better sense of what factors

predict teaching success. The reality is that very few of these mandated criteria are relevant to teaching quality or how pre-service teachers learn to teach (Steele, Pepper, Springer & Lockwood, 2015). Another significant limitation of the existing studies that have examined the role of personal factors during the preparation phase are that they often use data from a sample of very few teacher candidates during a single time point (see Driscoll & Pianta, 2010; Woolfolk, Rosoff & Hoy, 1990). These cross-sectional studies are useful in providing descriptive evidence of how personal factors influence teaching skills, but cannot identify relationships over time that inform program design. The lack of longitudinal data on how pre-service teachers' perceptions, beliefs and personalities influence their development of teaching skills precludes us from understanding how these relationships change over time and the implications this has for teacher preparation. If, for example, there was evidence that personal factors such as prior experience positively and consistently shaped how pre-service teachers learned to teach during the preparation phase, preparation programs might include more practice opportunities or targeted supports for candidates who did not enter the program with a significant amount of experience.

In Chapter 2, I examine how pre-service teachers' self-efficacy beliefs develop over the duration of a teacher preparation program. The lack of research examining trajectories of pre-service teachers' self-efficacy beliefs over the duration of the preparation phase is particularly concerning since there is evidence that these beliefs develop early on in a teacher's tenure (Bandura, 1977; Tschannen-Moran & Hoy, 2007) and persist across her career. In this chapter, I use linear growth models to explore trajectories of self-efficacy beliefs across the duration of the preparation program and

assess variation in these beliefs among 7 cohorts of individual teacher candidates (approximately 715 candidates) across three time points during the preparation phase. This is a substantially larger sample than has been previously used in studies on pre-service teachers' self-efficacy beliefs and the longitudinal nature of the data allow me to examine the development of these beliefs among the *same* teacher candidates, which is also rare in the literature.

Results suggest that although there is substantial variation in self-efficacy beliefs among individual teacher candidates, on average, their self-efficacy beliefs do not change significantly during the teacher preparation program. This is in contrast with previous work examining changes in self-efficacy beliefs at the start and end of preparation programs (see Dial, 2015; Fives, Hamman & Olivarez, 2007; Lin & Gorrell, 2001; Pendergast, Garvis & Keogh, 2011; Wenner 2001), but these studies use different samples of candidates at each time point and thus do not examine changes in self-efficacy beliefs among the *same* sample of candidates over time. I also examine whether teacher candidates' teaching specialty, their personality characteristics or attitudes towards teaching predict teacher candidates' trajectories of self-efficacy belief development. I find that teaching specialty (that is, whether teacher candidates were preparing to teach elementary, secondary or special education students) and their personality characteristics (levels of neuroticism and conscientiousness) predict change in self-efficacy beliefs during the teacher preparation phase.

Findings from this study provides empirical evidence that personal dispositions and choices of individual teacher candidates do affect their confidence in their ability to teach effectively in a classroom. Teacher education programs enroll a wide variety of

applicants who bring with them different characteristics, personalities and ideologies with the goal of training them to teach in stressful and demanding classroom environments. This study highlights the importance of understanding differences in candidates based on what they bring with them to the preparation program as well as the role these differences play in affecting their self-efficacy beliefs. Ultimately, helping teachers learn more about themselves and their beliefs may contribute to them developing more skills and confidence in their own ability as educators.

Chapter 3: Re-Examining the Evidence on Coaching in Teacher Education Using Systematic Replication

Along with moving some of the learning that happens "on-the-job" while teaching in a classroom to the preparation phase, a major concern among teacher educators is how to provide teacher candidates with standardized practice opportunities and scalable professional development that can be used to optimize candidates' instructional skills and preparedness (Kraft & Blazar, 2016). In particular, preparation programs across the country have been studying the degree to which some practice with targeted feedback can also occur in coursework. Such "approximations of teaching" -role-plays, rehearsals, and simulations-have been shown in qualitative work to support teacher candidates' ability to translate theoretical knowledge about "effective teaching" into their actual practice (Grossman et al., 2009a; Kavanagh & Rainey, 2017; Reisman et al., 2019; Windschitl, 2002).

Recent work has shown that coaching is a form of professional development and targeted feedback that improves in-service teachers' instructional skills, classroom organizational skills and attitudes towards teaching (Desimone & Pak, 2016; Kretlow &

Bartholomew, 2010). Although the literature on coaching in the pre-service context is nascent, a handful of studies associate coaching with improvements in teacher candidates' affect, overall satisfaction with preparation, attitudes towards self-development, and skill development (Bowman & McCormick, 2000; Cohen, Wong, Krishnamachari & Berlin, 2020; Cohen & Wiseman, 2021). However, because coaching is resource-intensive, teacher preparation needs more evidence on the best ways to support teacher candidates through expert coaching tied to standardized practice. Most of the existing research on coaching in the pre-service context is that most studies are small-scale, descriptive and do not allow for causal estimations of different kinds of supports in improving candidates' instructional skills (Kraft, Blazar & Hogan, 2018), particularly on the robustness of such coaching supports across systematic sources of variation (Hill, Beisiegel, & Jacob, 2013; Blazar & Kraft, 2015).

In Chapter 3, which is co-authored with Vivian Wong and Julie Cohen, we leverage mixed reality simulations to address the challenges with providing pre-service teacher candidates with standardized practice opportunities coupled with coaching during the preparation phase. In teacher preparation, mixed reality simulations can be designed to reflect the real-world complexities of live teaching environments (Slater, 2009), especially since studies have shown that simulations feel more realistic than other approximations of practice including role-plays and rehearsals, and that participants respond in ways that are closely aligned with what they might do in actual classroom settings (Arora et al., 2011a; Dieker, Rodriguez, Lignugaris/Kraft, Hynes & Hughes, 2013; Slater, 2009).

In this chapter, we use meta-analytic and replication methods to estimate the causal impact of providing pre-service teacher candidates with coaching, as well as understand the robustness of coaching effects across different systematic sources of variation that we might expect to find in teacher preparation. Chapter 3 presents results from a series of five experimental replication studies that evaluate the benefits of targeted, directive coaching on pre-service teachers' pedagogical skills in "mixed reality" simulation settings. We examine the pooled cross-study impact of coaching on participants' performance in the simulation setting, as well as the replicability of coaching effects across systematic sources of variation (including pedagogical task, timing of the study, context of practice and coaching and mode of delivery of the simulated session).

Findings from Chapter 3 show that providing participants with coaching significantly improves their teaching performance in a simulated teaching environment. Across the five experimental replication studies, improvements in overall quality of teaching as a result of receiving teaching ranges from 0.63 SD to 1.75 SD. The pooled meta-analytic effect across the six replication studies is 1.49 SD. Using distance-based correspondence measures (including magnitude of effects, direction of effects and statistical significance patterns) to identify patterns of robustness, we also find that coaching effects are robust across different teaching tasks, different study timing, as well as different modes of delivery of the simulated session, but are smaller in magnitude for a sample of participants who are not yet enrolled in formal teacher preparation compared to a sample of pre-service teacher candidates in a university-based teacher education program. Taken together, findings from this paper provides evidence on *the context and*

conditions under which targeted, directive coaching during standardized practice opportunities can be used to leverage improvement in teacher candidates' pedagogical skills. This evidence is critical in light of growing concern that traditional professional development programs fail to improve teaching practices and the need to identify specific conditions under which such programs produce better outcomes for teachers (Desimone, 2009).

Chapter 4: Evaluating the use of mental rehearsal techniques among pre-service teacher candidates during a simulated teaching opportunity

Novice teachers consistently report feeling underprepared and overwhelmed when they first begin to teach in a classroom environment, with researchers using terms such as "reality shock", "survival phase" (Huberman, 1989) and "transition shock" (Corcoran, 1981) to describe the challenging nature of the transition from pre-service to novice teacher. This is concerning because pre-service teachers are more vulnerable to the negative impacts of stress and uncertainty in the classroom environment, largely because of their inexperience in classroom settings, as well as an unclear perception of their own status as a teacher (Hopkins et al., 1997). There is also evidence to suggest that novice teachers are more susceptible to stress, burnout and attrition compared to their experienced peers (Roness, 2011).

This concern is particularly relevant in the context of classroom management, since it is an area that novice teachers report feeling the most ill-prepared to address (Latham & Vogt, 2007; Le Maistre & Paré, 2010; Sabar, 2004). Even when teacher candidates have the skills necessary to effectively redirect off-task student behavior, they often report feeling emotionally exhausted by the process of managing a classroom

environment (Jennings & Greenberg, 2009) on a regular basis. Research suggests that in order to effectively manage disruptive students, novice teachers need a combination of pedagogical knowledge regarding best practices as well as emotional resources to counter the stressful nature of redirecting student behaviors and managing a classroom. That is, when teachers are emotionally exhausted or perceive a classroom management task as a threat to their efficacy as a teacher rather than as an opportunity to learn, they are less likely to be effective at managing disruptive student behaviors (Seiz, Voss and Kunter, 2015).

One promising set of interventions that may be used to improve performance and perceptions related to classroom management for pre-service and novice teachers are mental rehearsals. Mental rehearsal is a cognitive reappraisal strategy that has been used in healthcare and sports settings to improve performance on a specific task related to the intervention, as well as participants' perceptions of the task including stress, self-efficacy and confidence levels (Arora eta al., 2011a; Eldred-Evans et al., 2013). Essentially, mental rehearsal uses a script (or stimulus) that focuses on strengthening the imagery representation of the skills that a participant is expected to use in a particular situation or task (Arora et al., 2011a). By allowing participants to mentally visualize themselves executing a task before performing it, mental rehearsal serves as a preparation technique that improves performance. In addition, the organized format of mental rehearsal scripts reduce cognitive load and benefits learners (particularly novices) by allowing them to cognitively reappraise the task in the absence of physical exercise, ultimately improving their perceptions of the task.

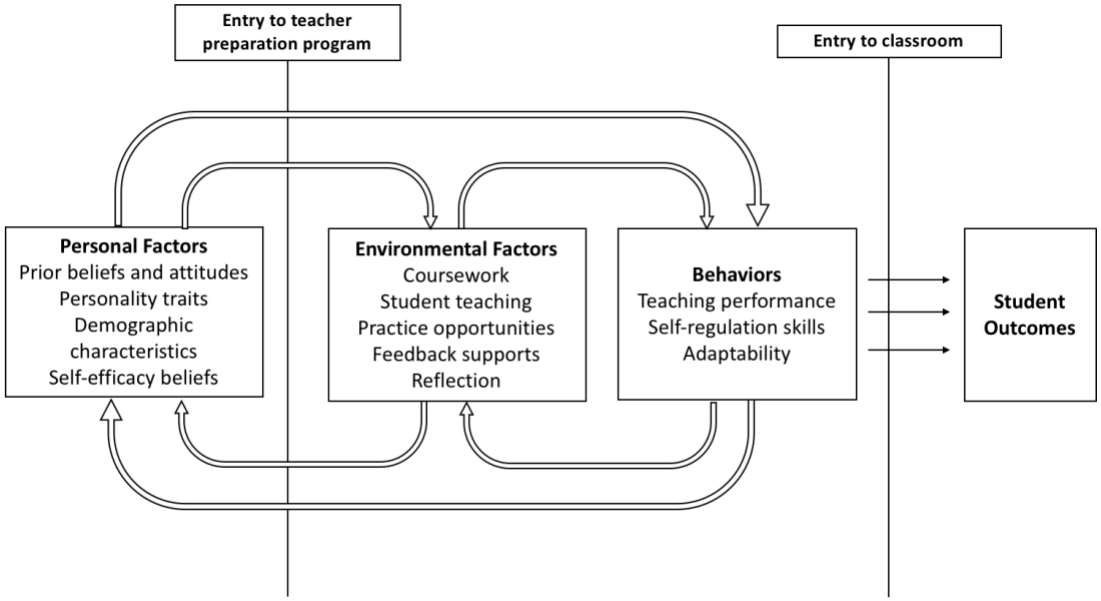
Chapter 4 addresses the gap in the literature on interventions that can help pre-service and novice teachers improve their performance and perceptions related to classroom management. In Chapter 4, I evaluate the efficacy of a cognitive appraisal technique called mental rehearsal at improving pre-service teacher candidates' performance and perceptions (including stress levels, self-efficacy beliefs, ratings of student avatar behaviors and endorsements of different management approaches) during a simulated teaching practice opportunity. The simulated teaching task involves redirecting off-task student behavior while setting classroom norms, a task that is often considered stressful by novice teachers (Le Maistre & Paré, 2010). Implemented as part of a larger experimental study on the efficacy of coaching supports on pre-service teacher candidates' instructional skills (explored in Chapter 3), a 2x2 factorial design allows me to understand the impact of mental rehearsal both as a stand-alone intervention as well as in combination with coaching.

Results suggest that mental rehearsal did not have a significant impact on candidates' stress levels or self-efficacy beliefs, but significantly improved their performance during the simulated teaching task. Specifically, I find that teacher candidates assigned to complete mental rehearsal provided more timely redirections to simulated student avatar behavior, more specific redirections, more succinct redirections, and scored higher on an overall measure of quality of redirections. In addition, I find that while mental rehearsal does not offer candidates who were randomly assigned to receive coaching in-between simulator sessions any advantage (over and above the benefit of receiving feedback from a coach), candidates who were randomly assigned to participate in self-reflection between simulator sessions benefited when they completed mental

rehearsal. Taken together, findings from this study provide important evidence on the use of a cognitive strategy such as mental rehearsal in facilitating the improvement of performance and perceptions among pre-service teacher candidates.

Figure 1

Teacher learning during the preparation phase



CHAPTER II: EXAMINING THE ROLE OF PERSONAL FACTORS IN THE DEVELOPMENT OF SELF-EFFICACY BELIEFS AMONG PRE-SERVICE TEACHERS

Teacher self-efficacy, or teachers' beliefs in their ability to promote student learning has been consistently linked to a variety of improved teacher and student outcomes. Teachers who report higher self-efficacy beliefs feel more confident in their ability to teach their students, adjust quicker to the classroom environment and are more confident in their ability to deliver effective instruction (Redman, 2015; Yost, 2006). In the classroom, efficacious teachers exhibit greater levels of planning, better classroom organization and improved teaching practices (Tschannen-Moran & Hoy, 2001; Zee & Koomen, 2016). Students in classrooms with teachers who have high self-efficacy beliefs demonstrate improved student self-efficacy (Anderson, Greene & Loewen, 1988; Lauermann & Berger, 2021), higher student motivation (Midgley, Feldlaufer & Eccles, 1989; Schiefele & Schaffner, 2015), higher levels of classroom engagement (Good & Brophy, 2003; Poulou, Reddy & Dudek, 2019) and increased academic achievement (Caprara, Barbaranelli, Steca & Malone, 2006; Zee & Koomen, 2016).

Although the literature establishes a clear link between higher self-efficacy beliefs and different teacher and student outcomes once the teacher enters the classroom, less is known about how and when self-efficacy beliefs are developed among pre-service teachers. Research suggests that self-efficacy beliefs form early on a teacher's career and once formed, these beliefs become resistant to change (Bandura, 1977; Tschannen-Moran

& Hoy, 2007). Pre-service teacher candidates enter preparation programs with some beliefs about their ability to teach, based largely on their prior experiences with and observations of teaching (Lortie, 1975; Richards & Lockhart, 1996). They also bring with them a set of personal characteristics, including their attitudes towards and beliefs about teaching, personality traits and demographic characteristics including age, gender, race/ethnicity, etc. that are likely to influence their self-efficacy beliefs. During the preparation experience, program components such as coursework, teaching experiences and mentor support interact with candidates' personal characteristics to change their existing self-efficacy beliefs. At this time, self-efficacy beliefs are also influenced by candidates' behaviors in response to their physical and social environment in the preparation program. Since we know that teacher self-efficacy beliefs are particularly malleable during the first few years of teaching (Pfitzner-Eden, 2016; Tschannen-Moran & Hoy, 2007), the interaction of personal factors, environmental factors and behavioral factors become critical to study during the preparation phase. Only by understanding how self-efficacy beliefs develop during the initial years of a teacher's tenure (i.e., preparation) and the role of personal factors in such development can we better equip teachers to be efficacious and effective in their classrooms from the beginning of their tenure.

Much of the literature on pre-service teachers' self-efficacy beliefs focuses on evaluating the role of Bandura's four theorized sources of these beliefs (mastery experiences, vicarious experiences, verbal persuasion and physiological arousal) rather than examining the predictors of teacher self-efficacy or the trajectories of their development throughout the preparation phase (Bautista & Boone, 2015; Brown, Lee &

Collins, 2015; Cantrell, Young and Moore, 2003; Wagler, 2011). Understanding predictors and trajectories of self-efficacy beliefs during the first few years of a prospective teacher's tenure is critical to understanding how confident teacher candidates are in their ability to effectively teach future students, as well as how to provide targeted supports during the preparation phase that can maximize such confidence. Of the studies that do examine the predictors of self-efficacy beliefs, most focus on international samples of pre-service teachers, particularly in Australia (see George, Richardson & Watt, 2018; Lazarides, Watt & Richardson, 2020; Ozder, 2011) and Turkey (see Kaner, 2010; Savran & Cakiroglu, 2003). Although helpful in understanding how self-efficacy beliefs are formed during the preparation phase, these studies offer limited evidence that may be applied to pre-service teachers or preparation programs in the American context since teacher candidates likely differ in important ways that influence their self-efficacy beliefs (including personality traits, attitudes towards teaching, preparation experiences and so on). A third and important limitation with the existing literature on self-efficacy beliefs is that these studies compare different cross-sectional samples of teacher candidates rather than longitudinally tracking the *same* cohort of teacher candidates over time. Thus, even though we know self-efficacy beliefs are critical to the success of a teacher and their students, it is difficult to ascertain when these beliefs are formed, how they change in the first few years of a teacher's tenure and whether they are influenced by personal characteristics of teacher candidates.

This study bridges gaps in the existing literature by identifying the role that specific personal characteristics including choice of teaching specialty (i.e., elementary, secondary, special education), personality traits and attitudes towards teaching play in the

development of self-efficacy beliefs during the preparation phase. The goals of this study are twofold: first, I aim to understand how pre-service teacher candidates' self-efficacy beliefs change during the preparation phase by examining their growth trajectories over time. However, the literature on self-efficacy beliefs suggests that examining aggregate patterns in self-efficacy beliefs among pre-service teachers may obscure important variation in how individual pre-service teachers' self-efficacy beliefs develop during the preparation period (Mitchel & King, 2016). To empirically address this issue, I then use these growth trajectories to explore variation in self-efficacy beliefs based on personal factors such as choice of teaching specialty, personality domains and attitudes towards teaching. Thus, the second goal of this study is to explore the role that these personal factors play in explaining variation in self-efficacy beliefs both at the beginning and across the duration of the preparation program.

In meeting these two goals, this paper makes important contributions to existing research on self-efficacy beliefs among pre-service teachers. First, this study leverages multiple years of longitudinal data on seven cohorts of pre-service teachers. Large-scale data collection, particularly longitudinal data tracking is rare in teacher preparation, therefore this study is unique in that it leverages data for multiple years for the *same* teacher candidates. These data allow for a comprehensive investigation of the role of personal characteristics in the development of pre-service teachers' self-efficacy beliefs across the duration of the preparation phase. The findings from this paper suggest that pre-service teacher candidates enter the preparation program with different self-efficacy beliefs based on their prior attitudes, personality traits and beliefs about teaching and that these factors significantly predict trajectories of self-efficacy development across the

duration of the program. Overall, my findings suggest that based on these different personal factors, individual teacher candidates likely derive different experiences from the same teacher preparation program as far as the development of their self-efficacy beliefs are concerned. The findings from this study have important implications for teacher preparation programs looking to provide targeted supports that enhance the preparation experience and improve candidates' beliefs in their ability to teach effectively in the classroom from day one.

Background

Self-efficacy Beliefs in Classroom Contexts

First described by Bandura using social cognitive theory, self-efficacy refers to an individual's beliefs about their capacity to produce designated levels of performance and influence outcomes and events affecting their lives (Bandura, 1977, 1986). Self-efficacy in the context of teaching and classrooms was first studied by Armor and colleagues who define teacher self-efficacy as "the extent to which the teacher believes he or she has the capacity to produce an effect on the learning of students" (Armor et al., 1976; pg. 31). In its simplest form, teacher self-efficacy can be thought of as a teacher's belief in their own ability to effectively teach their students in a manner that leads to improved student outcomes.

Teacher self-efficacy has always been of interest to researchers, primarily because efficacy beliefs are associated with improved teacher and student outcomes. The association between strong self-efficacy beliefs and improved teacher outcomes is particularly pronounced for novice teachers, who report greater satisfaction in teaching, less stress, and more favorable perceptions of how they were prepared for the classroom

(Hoy, 2000). Teachers with high self-efficacy also demonstrate higher levels of effort and persistence in their teaching (Skaalvik & Skaalvik, 2017), resulting in improved student outcomes including higher academic achievement (Caprara et al., 2006; Zee & Koomen, 2016), higher student motivation (Schiefele & Schaffner, 2015) and better attitudes towards school (Poulou et al.,, 2019). In essence, high teacher self-efficacy beliefs have the potential to improve teacher practices in the classrooms, teachers' perceptions about teaching more generally, and ultimately improve a number of student outcomes that are of policy interest.

Self-efficacy Beliefs During Teacher Preparation

In this study, I focus on self-efficacy beliefs among pre-service teachers enrolled in a teacher preparation program. Self-efficacy beliefs among pre-service teachers has gained more research interest following evidence that novice teachers enter the profession with high hopes regarding the impact they will be able to have, but often end up facing painful reality shocks (Tschannen-Moran & Hoy, 2007; Voss & Kunter, 2020). Teacher preparation programs across the nation are criticized for inadequately preparing new teachers, particularly in developing beliefs in their own ability to implement effective teaching and learning strategies (Zeichner, 2014; Riddle, 2018). In a study on the relationship between teachers' views of their preparedness and efficacy in classrooms, Darling-Hammond (2000) found that most teachers report feeling underprepared to teach and were prone to believing that students' learning was more influenced by peers and home environments than classrooms and teachers. Critics have long argued that traditional teacher programs provide candidates with content knowledge courses and pedagogical experiences, but short-change prospective teachers on raising self-

perceptions of teaching competence (Clark & Newbury, 2019; Tschannen-Moran, Hoy & Woolfolk-Hoy, 1998).

Although the literature suggests that preparation programs are responsible to some extent for inculcating self-efficacy beliefs among teacher candidates, it is important to note that pre-service teachers enter preparation programs with pre-formed self-efficacy beliefs. These beliefs are based on their previous learning experiences as students (Busch, 2010; Lortie. 1975), perceptions of teacher education (Da Silva, 2005; Li, 2020), experience of teaching, personality traits and broader educational principles (Richards & Lockhart, 1996). During the preparation program, self-efficacy beliefs are likely to be influenced by other factors, including formal training and the context of their teaching experiences (Clark & Newbury, 2019; Humphrey & Wechsler, 2006). Bandura (1997) suggests that teacher efficacy may be most malleable early in the learning phase, hence the role of preparation program and its components is central to the development of self-efficacy beliefs during the early years of a teacher's career. Thus, as candidates begin coursework and accumulate teaching experiences during the preparation phase, their self-efficacy beliefs are likely to evolve and change depending on how they cognitively process their own development as well as their reactions to their social and physical environments in the preparation phase (Bandura, 1997b; Tschannen-Moran & Hoy, 2007; Wyatt 2015). Figure 2, adapted from Bandura's Social Cognitive Theory, describes the triad of factors that influence how a pre-service teacher's self-efficacy beliefs are formed prior to, during and following their preparation experience.

Previous work on the development of self-efficacy beliefs among pre-service teachers has mostly focused on the four sources of self-efficacy theorized by Bandura

(mastery experiences, vicarious experiences, verbal persuasion and physiological arousal). In particular, there is a large body of literature examining the role of coursework and student teaching placements (that represent mastery experiences) in the development of self-efficacy beliefs, particularly in specific content areas (Cantrell et al., 2003; Gurvitch & Metzler, 2009; Pfitzner-Eden, 2016). However, limiting our understanding of how self-efficacy beliefs develop during teacher preparation by examining specific components of the program ignores other personal and behavioral factors that are integral to the development of these beliefs (Clark & Newberry, 2019). Teacher candidates' preparation and ability to teach effectively are shaped by the interaction of their personal background, their formal training (such as coursework) and the context of their student placements (Humphrey & Weschler, 2006). Therefore, personality traits and other individual factors are crucial to the development of self-efficacy beliefs, even as components of teacher preparation influence self-efficacy beliefs.

Personal factors in the development of self-efficacy beliefs

Traditionally, examining the role of personal factors in how pre-service teachers learn to teach has focused mostly on identifying which of these factors are predictive of teaching success in order to screen applicants to the teaching profession. Much of the existing research on pre-service teacher candidates' characteristics when entering preparation programs is limited to admission criteria including undergraduate GPA, SAT/ACT score and selectivity of undergraduate institution attended (Mitchel & King, 2016). It is surprising, therefore, that few of these mandated criteria are actually relevant to teaching quality (Steele et al., 2015). One explanation for the lack of association between traditional admission criteria and teaching success may be that focusing on these

discrete variables obscures the likelihood that it is a combination of variables including candidates' dispositions and traits in addition to the quality of the preparation program that makes an effective teacher (Mitchel & King, 2016), and not just criteria used to decide entry to the profession. Uncovering whether teacher candidates' demographic characteristics, personality traits or attitudes towards teaching predict both teaching success and teaching efficacy is important since it would provide preparation programs with another tool to identify, accept and train candidates who go on to become effective teachers.

Teacher preparation programs play a two-fold role in the development of self-efficacy beliefs (Wiens & Ruday, 2014). First, teacher preparation programs serve as gatekeepers into teaching in that they screen individuals who may or may not show potential to be effective teachers. To this end, examining differences in self-efficacy beliefs at the start of the preparation program based on personal factors can inform programs whether teacher candidates begin the preparation phase with different needs and thus require different supports. In addition, as emphasized throughout this study, it is important to note that personal factors may also interact with progress through the preparation experience to predict change in self-efficacy beliefs. Based on different characteristics and perspectives, individual teacher candidates may develop self-efficacy beliefs in different ways as they progress through program components including coursework and student teaching experiences.

As noted throughout the following sections, there are substantial limitations with the existing research on how personal factors such as choice of teaching specialty, personality traits and attitudes towards children affect self-efficacy beliefs during the

preparation phase. This paper significantly advances the existing literature on differences in self-efficacy beliefs among pre-service teachers based on their choice of teaching specialty by leveraging longitudinal data for multiple cohorts of teacher candidates enrolled in the *same* teacher preparation program. The longitudinal nature of the data allow me to explore differences in self-efficacy beliefs not just at the beginning (or end) of the program, but across multiple time points during the preparation program for the same candidates. Data for multiple cohorts allow me to examine trends in the development of self-efficacy beliefs across a larger sample of pre-service teacher candidates, and the systematic data collection by a single university-based teacher preparation program allows me to examine differences due to the three personal factors explored in this study (choice of teaching specialty, personality traits and attitudes towards children and teaching), and not because of differences in programs.

Choice of teaching specialty and self-efficacy beliefs. When enrolling in university-based teacher preparation programs such as the one examined in this study, pre-service teacher candidates choose to teach elementary students, secondary students or students with disabilities. According to the literature, this choice is largely based on prior perceptions of teaching that pre-service teacher candidates bring with them to the program or on motivational factors that are closely tied with their own educational experiences (Hong et al., 2017). Commonly referred to as teaching specialty, literature on teacher preparation has documented that pre-service elementary teachers differ from pre-service secondary teachers in their attitudes, perceptions of classroom problems (Ponnock, Torsney & Lombardi, 2018; Veenman, 1984) and commitment to teaching (Evans and Tribble, 1986; Ponnock et al., 2018).

Although the current literature comparing self-efficacy beliefs between elementary, secondary and special education teachers is limited, it provides evidence that teaching specialty and self-efficacy beliefs may be associated. An early study by Evans and Tribble (1986) descriptively examined differences in perceived problems and self-efficacy beliefs among 179 elementary and secondary pre-service teachers in the first year of their preparation programs. The authors find that prospective elementary teachers reported significantly higher self-efficacy beliefs than prospective secondary teachers ($p<0.01$). This finding is in direct contrast to another study by Savran and Cakiroglu (2003). In their paper, the authors examined differences in science efficacy beliefs among 646 elementary and secondary pre-service teachers in Turkey and found that secondary teachers reported significantly higher science self-efficacy beliefs than their elementary counterparts at the end of the preparation phase.

Another study by Freytag (2001) examined differences in self-efficacy beliefs among 48 (36 elementary/secondary and 12 special education) beginning teachers who had just completed their teacher preparation programs. Freytag found that special education teachers reported significantly higher self-efficacy beliefs than their general education elementary/secondary peers. A similar study in Turkey with a sample 234 teachers also reported divergent findings, in that participants who taught students with special education needs did not report significantly different self-efficacy beliefs than their peers who taught students in general elementary and secondary classrooms (Kaner, 2010).

Although helpful in exploring the relationship between choice of teaching specialty and self-efficacy beliefs, these studies are limited by their cross-sectional design

(Evans and Tribble examined differences in self-efficacy at the beginning of the preparation phase, whereas Savran and Cakiroglu examined differences at the end of the preparation period). In addition, Freytag (2001) examined differences across beginning teachers who enrolled in different preparation programs, so the programs themselves (and not just specific components) may also explain variation in self-efficacy beliefs.

Personality traits and self-efficacy beliefs. In the past several decades, teacher education has consistently attempted to measure and identify personality traits that teacher preparation programs might view as necessary for their candidates to possess to become effective teachers (Bolding, 2017; Damon, 2007; Sockett, 2009). The literature consistently points to the importance of personality characteristics in determining how individuals perceive their skills and abilities, particularly when they are inexperienced to the extent that pre-service teacher candidates are during the preparation phase (Jamil, Downer & Pianta, 2012). A popular framework used by a large number of research studies to identify personality traits is the "Big Five" conceptualization of personality developed by Costa and McCrae (1992), which measures an individual's neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Ripski, LoCasale-Crouch and Decker (2011) characterize neuroticism as negative emotions such anxiety and low self-esteem, extraversion as being sociable and assertive, openness as being curious and imaginative, agreeableness as sympathetic and easily moved and conscientiousness as being responsible and determined.

Existing research suggests that on average, pre-service teachers report higher levels of all five personality traits compared to the general population i.e., pre-service teachers report higher neuroticism, higher extraversion, higher openness, higher

agreeableness and higher conscientiousness (Decker & Rimm-Kaufman, 2008). Although we might expect personality traits to remain stable across the pre-service and in-service phases, Diez (2007) believes that teacher preparation programs play an important role in supporting candidates to adapt their personalities in ways that set them up for success in the classroom. As they learn to teach, it is likely that pre-service teachers mediate their personal orientations in relation to professional contexts and practices, influencing their daily decisions and judgments in the classroom context. Importantly, teacher preparation programs are key in enabling teacher candidates to transcend personal preferences and achieve dispositions that allow them to teach effectively in classrooms (Carroll, 2012).

Thus far, only a handful of studies have examined whether different personality domains predict self-efficacy beliefs in teaching contexts. Using the same five-factor model employed in this paper, a study by Navidnia (2009) found that extraversion and conscientiousness were significantly associated with self-efficacy beliefs among a sample of 168 English language teachers. Specifically, teachers who expressed higher levels of extraversion reported lower levels of self-efficacy beliefs, and candidates who expressed higher levels of conscientiousness reported higher self-efficacy beliefs. Another paper by Djigic, Stojiljković and Dosković (2014) examined whether 168 in-service teachers' self-efficacy beliefs were connected to basic personality dimensions. The authors found that while all five personality domains significantly predicted self-efficacy beliefs among participants, conscientiousness and openness were the most significant predictors. Both domains were positively related to self-efficacy beliefs, with participants who were more open and more conscientious reporting higher self-efficacy beliefs. Findings from both papers are consistent with a more recent study on pre-service teacher candidates by Jamil

and colleagues (2012). In their paper, the authors examined a sample of 509 pre-service teacher candidates across 4 cohorts of a teacher preparation program and found that pre-service teachers' levels of extraversion and neuroticism were predictors of their self-efficacy levels at the end of the program. Teacher candidates who were more extraverted reported higher levels of self-efficacy, while teacher candidates who were more neurotic reported lower levels of self-efficacy.

Although all three studies are informative on the link between personality domains and self-efficacy beliefs, both the Navidnia (2009) and Djigic et al. (2014) studies use samples of in-service teachers rather than pre-service teachers. If, as pointed out by Carroll (2012), teacher preparation programs are uniquely situated to providing pre-service teacher candidates with different supports to enable them to teach effectively, understanding the link between personality traits and self-efficacy beliefs during the pre-service phase is crucial. Another limitation with the studies highlighted above is that they use cross-sectional samples of teachers/ teacher candidates to understand the relationship between personality traits and self-efficacy beliefs at a specific time point rather than over the duration of the preparation experience. As highlighted by the Jamil et al. (2012) paper, more longitudinal research that verifies the relationship between personality factors and the self-efficacy is critical to our understanding of the development of such beliefs.

Attitudes towards teaching and self-efficacy. A third set of variables that likely influence the development of self-efficacy beliefs among pre-service teachers are their attitudes towards teaching and students. Early research on teacher education suggests that candidates enter teacher preparation programs with pre-conceived attitudes about

teaching, children and classrooms based on their own experiences as students (Kagan, 1992; Lortie, 1975; Pajares, 1992; Decker & Rimm-Kaufman, 2008). These attitudes shape their teaching practices, at least for the first few years they spend in a classroom (Pianta, 2005). More broadly, attitudes towards children can be classified into teacher-centered and student-centered approaches. Teacher-centered approaches are found where judgements about appropriate teaching practices, legitimacy of information and the way knowledge is operationalized rest primarily with the teacher (Kain, 2003; Rubin & Herbert, 1998; Villaume, 2000). In contrast, student-centered approaches are when the construction of knowledge is a shared process and is largely achieved through student engagement in classroom activities.

There is hardly any literature examining the link between pre-service teacher attitudes and their self-efficacy beliefs. In their paper, Jamil et al. (2012) examine the relationship between attitudes towards children and self-efficacy beliefs, and find that pre-service teachers who had higher self-efficacy beliefs adopted more child or student-centered approaches than pre-service teachers who reported lower self-efficacy beliefs. Although indicative of how attitudes towards children and self-efficacy beliefs are linked, one major limitation with the paper by Jamil and colleagues is that the authors use data on predictors and self-efficacy beliefs collected at the end of the teacher preparation. Thus, although the findings suggest that attitudes towards teaching and children do play a role in determining self-efficacy beliefs, we do not have evidence on whether this relationship holds at other time points during the teacher preparation phase, or how the relationship between attitudes towards teaching and self-efficacy beliefs change during the preparation phase. Examining this relationship is particularly important if we expect

that attitudes about teaching are likely to influence the development of self-efficacy beliefs during the preparation phase.

Research Methods

In this study, I leverage archived longitudinal data collected by a southern school of education in a large public university on prospective teacher candidates enrolled in an undergraduate/graduate combined teacher preparation program. The longitudinal nature of the data allow me to observe self-efficacy beliefs for the *same* teacher candidates over time, and multiple cohorts allow me to observe patterns in the development of self-efficacy beliefs for a *much larger sample* of pre-service teacher candidates than previous studied. Between 2009 and 2016, seven cohorts of teacher candidates completed self-report surveys about their education experiences, their attitudes towards teaching and working with children as well as a series of inventories including a Big Five personality measure. Candidates also completed self-report surveys on their self-efficacy beliefs at multiple time points during their enrollment: once each at the beginning of the program, approximately midway through the program, and at the end of the program. In my analyses, I include teacher candidates who have at least one time point of data, resulting in approximately 1500 observations and around 716 individual teacher candidates. This is a substantially larger sample than has been previously used in studies on pre-service teachers' self-efficacy beliefs and the longitudinal nature of the data allow me to examine the development of these beliefs among the *same* teacher candidates, which is also rare in the literature.

Setting

All participants in this sample were enrolled in a five-years Bachelors/Masters in Teaching (B/MT) teacher preparation program. In the first two years of the program, candidates completed general coursework that is required of all undergraduate students at the institution. Beginning in their third year, teacher candidates were officially enrolled in the teacher preparation program (which I operationalize as the beginning of the preparation program). Teacher candidates began their first year in the preparation program by participating in a year-long tutoring experience with a student from the local school district classified as struggling in addition to coursework on general and content-specific teaching, assessment techniques and classroom management among others. The tutoring experience is designed to allow candidates to begin working with students in a structured and focused manner where teacher candidates receive detailed feedback from faculty in the teacher education program. In their second year, teacher candidates completed two additional experiences teaching approximately 30 hours in each semester, prior to when they engaged in a formal student teaching opportunity in their third and final year. The structure of the specific program in this study is different from many preparation programs that encourage students to engage in teaching experiences, both informal and formal, only after they have completed foundational coursework. An important implication of this unique practice-based design allows me to examine changes in self-efficacy beliefs among teacher candidates who are provided with regular opportunities to practice their teaching skills.

Measures

This study uses a number of measures collected by the teacher preparation program through self-report surveys on teacher candidates' demographic information,

their personality traits and ideas about teaching in addition to administrative data on candidates' teaching specialty.

Demographic and administrative data. The teacher preparation program collected survey data on teacher candidates' demographic information, including their age, gender, race/ethnicity, parental education, characteristics of their high school (average SES status, average achievement levels and majority student race). These data were used as control variables in my analyses. The program also provided administrative data including teacher candidates' teaching specialty – i.e., whether they were preparing to teach elementary students, secondary students or students with disabilities. A summary of participant characteristics is provided in Table 1.

Personality factors. Personality factors were assessed using the Neo Five-Factor Inventory (Costa & McCrae, 1992). The NEO inventory is a 60-item version designed to provide a reliable and valid measure of participants' neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Responses range from 1 to 5 on a Likert scale: 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree and 5=Strongly Agree. As designed by the instrument's authors, the 60 items can be divided into five subscales representing neuroticism, extraversion, openness to experience, agreeableness and conscientiousness. Alphas on the NEO dimensions ranged from 0.77 to 0.87. Means and standard deviations are reported in Table 2.

Ideas about children. In this paper, I use the Ideas About Children scale (sometimes referred to as the Modernity scale) developed by Shaefer and Edgerton (1985) to differentiate between adult-centered views and child-centered views of teaching using a 16-item scale. Although developed originally to characterize parental views, the

scale has been used to characterize novice teacher approaches in many studies (Jamil et al., 2012; Mashburn, Hamre, Downer & Pianta, 2006). Each item is on a 5-point Likert scale rating from 1-strongly disagree to 5- strongly agree and the 16 items are averaged to create a single composite score. Lower scores indicate a more child-centered view and higher scores reflect a more adult-centered view. Items include statements such as "Children should always obey the teacher" and "In order to be fair, a teacher must treat all children alike". Alphas ranged from 0.72 to 0.75. Means and standard deviations are reported in Table 2.

Teacher self-efficacy beliefs. The main outcomes of interest in my analyses (pre-service teacher candidates' self-efficacy beliefs) were collected using the Teacher Sense of Efficacy Scale (TSES) developed by Tschannen-Moran and Hoy in 2001. The long-form version of the instrument consists of 24 items that are classified into three composite subscales: *1) efficacy for instructional strategies, 2) efficacy for student engagement and 3) efficacy for classroom management*. While the TSES scale was originally developed for use with in-service teachers, it has been used extensively with the pre-service teacher population as well. As recommended by the authors when using the instrument with pre-service teachers, I use scores on the overall scale rather than the subscales (Tschannen-Moran & Hoy, 2001)[1]. Responses range from 1-9 on a Likert scale with anchors at 1=Nothing, 3=Very little, 5=Some influence, 7=Quite a bit and 9=A great deal. Cronbach's alpha for the three subscales ranged from 0.88 to 0.90. Table 3 presents average self-efficacy levels and standard deviation for the overall self-efficacy scale at

[1] I also conducted factor analyses and tests of measurement invariance across time and cohort to ensure that the measure was invariant. Consistent with Tschannen-Moran and Hoy's (2007) factor analysis on the TSES measure for pre-service teachers, I find that the factor structure for the three subscales resulted in a single factor (for all time points and all cohorts) suggesting that the TSES measures the underlying construct of self-efficacy and that a total score is appropriate for the analyses presented in this paper

three time points for the sample of teacher candidates in this study. All teacher candidates completed the TSES measure as part of a programmatic survey that they were required to take at the beginning of the preparation program, in their second year of the program, and in their final year.

Analytic Model

I begin my analysis by examining whether teacher candidates' self-efficacy beliefs change significantly over the duration of the teacher preparation program. I then explore the variation in teacher candidates' self-efficacy beliefs at the beginning of the preparation program as well as in their growth trajectories over time. To do so, I fit a linear growth model to describe the shape of teacher candidates' trajectories using hierarchical linear modeling (Raudenbush, Bryk, Cheong, Congdon & Du Toit, 2000) in STATA. For each teacher candidate, a linear regression model is fit with time as the sole predictor (referred to as an unconditional model). The first part (Level 1) of the two-level hierarchical model examines variation in self-efficacy beliefs across time within each teacher candidate and is presented in Equation 1 below:

$$Y_{ti} = \pi_{oi} + \pi_{1i}(Time)_{ti} + e_{ti} \quad (1)$$

In Equation 1, each teacher candidate *i*'s self-efficacy beliefs (*Y*) at time *t* is a function of an intercept parameter π_{oi} (which represents their self-efficacy beliefs at the start of the teacher preparation program), growth parameter π_{1i} which represents the average change in self-efficacy beliefs over the duration of the teacher preparation program and the residual error term e_{ti} which represents the portion of teacher candidate *i*'s self-efficacy belief at time *t* that is not predicted by time. Here, I assume that the error term e_{ti} is independent and is normally distributed with a variance of σ^2.

At level 2, both the intercept and growth parameters are allowed to vary as a function of teacher candidate characteristics (i.e., each teacher candidate is allowed to have an individual intercept and trajectory). Thus, the level 2 model allows me to examine variation in self-efficacy beliefs between individual teacher candidates which may be otherwise be masked when examining aggregate results from the level 1 model. In the level 2 model presented in Equation 2 below, there are two level-2 random effects u_{0i} and u_{1i} with variances T_{00} and T_{11} respectively, and with covariance T_{01}.

$$\pi_{oi} = \beta_{00} + u_{0i}$$

$$\pi_{1i} = \beta_{10} + u_{1i} \quad (2)$$

Both the level 1 and level 2 models can be combined into the model presented below in Equation 3:

$$Y_{ti} = [\beta_{00} + \beta_{10} Time_{ti}] + [u_{0i} + u_{1i} Time_{ti} + e_{ti}] \quad (3)$$

Change in self-efficacy beliefs. In equation (3), the mean intercept β_{00} represents the average self-efficacy beliefs that teacher candidates reported at the beginning of the program, and the mean growth rate β_{10} represents the average change in self-efficacy beliefs between each time point (beginning to middle, middle to end of program). A positive, significant coefficient for β_{10} suggests that teacher candidates' self-efficacy beliefs are growing over time, and that the difference in self-efficacy beliefs between any two points is statistically significant.

Individual differences among teacher candidates. Equation (3) also includes estimates of variance for both the mean intercept and the mean growth rate. u_{0i} represents the variance in the mean intercept and provides a sense of the variability in self-efficacy beliefs at the beginning of the program. A large, statistically significant

estimate for u_{0i} suggests that teacher candidates' self-efficacy beliefs vary greatly at the beginning of the program, and that this variation is significant. Similarly, $u_{1i}Time_{ti}$ represents individual differences in how self-efficacy beliefs develop over time. A large, statistically significant estimate for $u_{1i}Time_{ti}$ suggests that teacher candidates' self-efficacy beliefs develop differently over time and highlights the need to examine *why* different teacher candidates' self-efficacy beliefs follow different trajectories.

The role of personal factors in the development self-efficacy beliefs. After characterizing individual differences in self-efficacy beliefs, I assess whether these beliefs are predicted by teacher candidates' teaching specialty, their personality domains (neuroticism, extraversion, openness, agreeableness and conscientiousness) and their attitudes about children and teaching. I also assess whether these factors predict changes in self-efficacy beliefs- that is, whether teacher candidates' teaching specialty, their personality domains and their attitudes predict changes in self-efficacy beliefs. To do so, I use a hierarchical linear growth model that includes level 2 predictors. The level 1 and level 2 models are outlined in Equations 4 and 5 respectively:

$$Y_{ti} = \pi_{oi} + \pi_{1i}(Time)_{ti} + e_{ti} \tag{4}$$

$$\begin{aligned}\pi_{oi} &= \beta_{00} + \beta_{01}TeachingSpecialty_i + \beta_{02}PersonalityDomain_i + \\ &\quad \beta_{03}Attitudes_i + u_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11}TeachingSpecialty_i + \beta_{12}PersonalityDomain_i + \\ &\quad \beta_{13}Attitudes_i + u_{1i}\end{aligned} \tag{5}$$

In the level 1 and level 2 models outlined above, the interpretations for the intercept and growth parameters remains the same from the unconditional growth model presented in Equations 1 and 2. These models can be combined as presented in Equation 6 below:

$$
\begin{aligned}
Y_{ti} = [&\beta_{00} + \beta_{10}Time_{ti} + \beta_{01}TeachingSpecialty_{ti} \\
&+ \beta_{11}(TeachingSpecialty)(Time)_{ti} + \beta_{02}PersonalityDomain_{ti} \\
&+ \beta_{12}(PersonalityDomain)(Time)_{ti} + \beta_{03}Attitudes_{i} \\
&+ \beta_{13}(Attitudes)(Time)_{ti}\,] \\
&+ [u_{0i} + u_{1i}Time_{ti} \\
&+ e_{ti}] \qquad (6)
\end{aligned}
$$

In the model above, β_{01} represents the magnitude of the relationship between teaching specialty and self-efficacy beliefs for each teacher candidate *i* at time *t* after controlling for other predictors. Similarly, β_{02} and β_{03} represent the magnitude of the relationship between personality domains and self-efficacy beliefs and attitudes towards children and self-efficacy beliefs respectively. β_{11} represents the magnitude of the relationship between teaching and self-efficacy beliefs over the duration of the teacher preparation program after controlling for other predictors. Similarly, β_{12} and β_{13} measure the relationship between personality domains and self-efficacy and teaching specialty and self-efficacy over the duration of the program respectively. All the other terms are interpreted as in Equation 3.

Results

Descriptive Statistics.

Table 1 describes the sample of pre-service teacher candidates included in my final analytic sample. Across the seven cohorts of candidates, 55%[2] of participants were female and a majority of participants were below the age of 24 (86%). In terms of racial/ethnic background, 80% of the sample reported being Caucasian, followed by Asian (10%) and African American (5%). A slight majority were preparing to teach secondary students (46%) and 41% were preparing to be elementary teachers. The remaining teachers were preparing to teach elementary students with special education needs (13%). Overall, the demographics of the samples used in this study are similar to those of individuals entering the teaching profession (American Association of Colleges for Teacher Education, 2013).

Change in self-efficacy beliefs

Next, I examine whether pre-service teacher candidates' self-efficacy beliefs changed significantly over the duration of the preparation program. Column 2 in Table 4 reports the results from the combined model presented in Equation (3). The estimated mean intercept, β_{00} , and mean growth rate, β_{10}, for the self-efficacy data were 7.23 and 0.01 respectively. These coefficients suggest that teacher candidates started the teacher preparation program with average self-efficacy beliefs around 7.23 points (on a 9-point scale) and that candidates' beliefs changed by an average of 0.01 points between each time point during the study. Starting the preparation program with average self-efficacy beliefs of around 7.23 points suggest that pre-service teacher candidate entered with relatively high confidence in their ability to teach students. This aligns with other studies

[2] Data on participants gender are missing for 32% of the sample because of administrative changes in how these data were collected and reported. On average, the teacher preparation program examined in this paper enrolls about 86% female and 14% male teacher candidates.

that have measured self-efficacy beliefs at the beginning of the teacher preparation program using the TSES measure (see Fives et al., 2007; Pendergast et al., 2011). The growth rate suggests that on average, teacher candidates' beliefs did not change significantly over the course of the teacher preparation program ($p>0.10$). Although surprising, this lack of overall change in self-efficacy beliefs has been documented by other studies including Lin and Gorrell (2001) who suggest that there is a need to examine variation among individual pre-service teachers rather than assess average self-efficacy beliefs across all pre-service teachers.

Individual differences among teacher candidates

Consistent with the theory that aggregate patterns of self-efficacy beliefs may obscure variation in individual self-efficacy beliefs, the HLM analyses show that there is substantial variation in self-efficacy beliefs among teacher candidates at this university. Results from the model presented in Equation (3) include estimates of variance of change that represents individual differences in how self-efficacy beliefs develop over time. Standard deviations (square roots of these variances) are presented in the random effects section of Table 4. The estimates for the variances of initial self-efficacy beliefs and change rate are 0.81 and 0.23 respectively. Both estimates are statistically significant, suggesting that teacher candidates vary significantly in terms of self-efficacy beliefs when they enter the program, and that there are also individual differences among the rate of change in teacher candidates' self-efficacy beliefs.

The role of personal factors in the development self-efficacy beliefs

I then examine whether teacher candidates' teaching specialty, personality traits or beliefs about teaching explain the variation in self-efficacy beliefs both at the beginning

of the program and in the trajectory of beliefs over the duration of the preparation program. Models 2-4 in Table 4 report estimates from the model presented in Equation 6 for each of the predictors. Model 1 in Table 4 controls for demographic factors, including teacher candidates' age, gender and race as well as fixed-effects for cohort (γ_j) and year of TSES measure completion. Each successive model adds teaching specialty, personality domains and ideas about children respectively.

Teaching Specialty. Model 2 in Table 4 estimates the relationship between teaching specialty and teacher candidates' self-efficacy beliefs after controlling for demographic variables, year and cohort fixed-effects. Estimates from this model suggest that elementary teacher candidates begin the preparation program with an average self-efficacy belief of 7.30 points and that their beliefs decrease by 0.02 points (on a 10-point scale) on average between each time point. The difference between elementary and secondary candidates is statistically significant (β_{01}=-0.30, p<0.05) suggesting that secondary teacher candidates report significantly lower self-efficacy beliefs on entering the program than elementary teacher candidates. Teacher candidates preparing to teach students with special education needs also report slightly lower self-efficacy beliefs than elementary teacher candidates at the start of the preparation program, but this difference is small in magnitude and is not statistically significant (β_{01}=-0.01, p>0.10).

Random effects estimates for model 2 suggests that there are significant differences in both the average self-efficacy beliefs at the beginning of the preparation program (β_{11}=0.74, p<0.01) and in the trajectory of self-efficacy beliefs based on teaching specialty β_{11}=0.22, p<0.01). Figure 3 plots the trajectories of self-efficacy beliefs by teaching specialty. As highlighted above, secondary teacher candidates begin

the preparation program with significantly lower self-efficacy beliefs than their elementary and special education peers. Across the duration of the preparation program, Figure 3 shows that for both elementary and special education teacher candidates, self-efficacy beliefs decrease slightly between the beginning and middle of the program, and then increase again by the end of the program. In contrast, Figure 3 shows that self-efficacy beliefs for secondary teacher candidates increase slightly between the beginning and middle of the program, and then decrease by the end of the program. Importantly, secondary teacher candidates continue to report lower self-efficacy beliefs than their peers throughout the program. It is important to note here that although I observe slight increases and decreases in self-efficacy beliefs across the duration of the program, on the whole, self-efficacy beliefs are remarkably stable across time for the sample of teacher candidates observed in this study.

Personality traits. Model 3 of Table 4 presents HLM estimates of different personality domains on teacher candidates' self-efficacy beliefs after controlling for teaching specialty, demographic variables and fixed-effects for cohort and year of completion. Estimates from model 3 suggest that after accounting for the other personality domains, only neuroticism and conscientiousness appear to significantly predict teacher candidates' self-efficacy beliefs at the beginning of the teacher preparation program. Specifically, teacher candidates who report higher levels of neuroticism at the start of the program report significantly lower self-efficacy beliefs than their peers who report lower levels of neuroticism at the start of the program (β_{01}=-0.26, p<0.10). The left panel of Figure 4 plots the relationship between different levels of neuroticism and self-efficacy beliefs at the three time points during the teacher

preparation program. Figure 4 shows that the relationship between self-reported levels of neuroticism and self-efficacy beliefs also appear to be stable across the duration of the preparation program. The stability of patterns in Figure 4 suggests that teacher candidates who report the highest levels of neuroticism consistently report the lowest levels of self-efficacy beliefs (and vice versa).

Apart from neuroticism, conscientiousness also appears to be significantly associated with pre-service teacher candidates' self-efficacy beliefs. Estimates from model 3 suggest that teacher candidates who report higher levels of conscientiousness at the start of the program report significantly higher self-efficacy beliefs than their peers who report lower levels of conscientiousness at the start of the program (β_{01}=0.92, p<0.01). Despite reporting higher self-efficacy, highly conscientious teacher candidates experience significant decreases in self-efficacy beliefs across the duration of the teacher preparation program (β_{11}=-0.20, p<0.05). This relationship is plotted in the right panel of Figure 4- of the five levels, teacher candidates who report very high conscientiousness report the highest self-efficacy beliefs at the start of the teacher preparation program but also experience the steepest decline.

Attitudes towards teaching. Finally, Model 4 of Table 4 presents the effect of attitudes towards teaching (measured by the Ideas about Children scale) on teacher candidates' self-efficacy beliefs after controlling for demographic factors, teaching specialty, personality domains and fixed effects for cohort and year of completion. HLM estimates suggest that teacher candidates who report more child-centered views of teaching at the beginning of the teacher preparation program report significantly higher self-efficacy beliefs (β_{01}=0.46, p<0.01) than their peers who report more adult-centered

views towards teaching. However, estimates of β_{11} suggest that attitudes towards children and teaching do not predict change in self-efficacy beliefs over the duration of the preparation program (β_{11}=0.00, p>0.10).

Discussion

Although personal characteristics and the manner in which they influence how an individual interacts and responds with the environment are critical to the development of self-efficacy beliefs according to Bandura's social cognitive theory, the teacher preparation literature has not paid much attention to the role of these factors in self-efficacy beliefs among pre-service teachers. Acknowledging and examining variation in the how individual pre-service teachers' self-efficacy beliefs is critical to tailoring a package of coursework, clinical practice, mentoring, and appropriate placement to fit the needs of different individuals. As they learn to teach, pre-service teachers need help mediating their personal beliefs, attitudes and ideas in relation to professional contexts and practices (Carroll, 2012). Essentially, candidates need assistance in building on the foundation of values, personal beliefs and personality domains they bring to teacher education to construct a responsible professional identity and set of practices.

This study presents important evidence on the role of personal factors including teaching specialty, personality traits and attitudes towards children in the development of self-efficacy beliefs among pre-service teacher candidates using data that are rare in typical teacher preparation contexts. I leverage rich longitudinal data collected by a single university-based teacher preparation program that allows me to observe teacher candidates' personal factors as well as self-efficacy beliefs across multiple time points during the preparation experience for multiple cohorts of the same teacher candidates.

These data allow me to improve and extent on previous cross-sectional studies examining differences and predictors of self-efficacy beliefs among pre-service teachers.

Overall, I find that when candidates enter their teacher preparation program several personal factors significantly predict their self-efficacy beliefs. First, I find that choice of teaching specialty predicts teacher candidates' self-efficacy at the start of the preparation program. Specifically, teacher candidates who are preparing to teach secondary students report significantly lower self-efficacy beliefs at the beginning of the preparation program as compared to teacher candidates who are preparing to teach elementary students or students who require special education. This difference is in line with earlier findings by Evans and Tribble (1986) who also documented significant differences among elementary and secondary teacher candidates. This paper improves on earlier literature by highlighting that differences between secondary candidates' self-efficacy beliefs and those of their peers continue to persist across the duration of the preparation program. These persistent differences suggest that secondary teacher candidates likely respond differently to the preparation program than elementary or special education candidates. For example, elementary and/or special education teacher candidates may benefit more from coursework and practice opportunities that help them meet children's needs for structure, discipline and routine, while secondary teacher candidates may benefit more from addressing the developmental needs of adolescents (Decker & Rimm-Kaufman, 2008). Tailoring coursework and practice opportunities to the specific demands that differ across different teaching specialties may be one way to improve teacher candidates' self-efficacy beliefs.

Second, I find that candidates' personality traits (neuroticism, extraversion, agreeableness and conscientiousness) also affect their self-efficacy beliefs at the beginning of the program, and that neuroticism and conscientiousness continue to affect self-efficacy beliefs throughout the program. These findings are consistent with those from Jamil et al. (2012) who find that teacher candidates who were more extraverted reported higher levels of self-efficacy, while teacher candidates who were more neurotic reported lower levels of self-efficacy. The ability to empirically link teacher candidates' personality domains and self-efficacy beliefs is critical to tailoring individualized experiences that allow candidates to transcend their personal dispositions and become effective teachers. Teacher education programs educate individuals from a wide range of backgrounds, experiences and education levels, all of which affect their teaching practice. Although personality characteristics are generally thought of as being "traits" and therefore hard to change, teacher education programs can help candidates identify the relationship between their own personalities and their teaching strategies as well as provide them with appropriate coping mechanisms to channel their personalities into effective teaching (Rimm-Kaufman & Hamre, 2010; Jamil et al., 2012). To provide an example, Rimm-Kaufman and Hamre (2010) highlight that an extroverted teacher candidate may need support from mentor teachers or coaches when presented with a challenging group of students and may fail if such support were not provided to them. In contrast, an introverted teacher may have developed strategies for their own teaching that is less dependent on external sources of support. Recognizing and addressing these different needs early on in the preparation phase might help teacher candidates realize

their own strengths and limitations accordingly, as well as enable preparation programs to provide differential supports that maximize pre-service teachers' efficiencies.

Finally, I find that attitudes towards children also explain a significant amount of variation in pre-service teachers' self-efficacy beliefs at the beginning of the program, but do not predict the trajectory of self-efficacy beliefs across the duration of the program. Specifically, I find that teacher candidates who express more child-centered approaches to teaching report significantly higher self-efficacy beliefs than those with more adult-centered approaches to teaching. These patterns are consistent with previous literature suggesting that in-service teachers with more child-centered approaches to teaching report higher self-efficacy beliefs than their peers who report more adult-centered approaches to teaching (Jamil et al., 2012). More research on the stability of attitudes towards teaching during the teacher preparation phase and factors that influence whether pre-service teachers adopt more child-centered views of teaching or adult-centered views would help uncover if we might expect to see a strong relationship with self-efficacy beliefs over time.

Limitations and Future Research

Although this study builds on previous work examining self-efficacy beliefs among pre-service teachers in important ways, it is limited by the fact that the data comes from a single teacher preparation program and thus these results might not generalize to other pre-service teaching populations. This study is also inherently descriptive in nature and makes no causal claims about the relationships between any personal factors studied and teacher candidates' self-efficacy beliefs. Findings from this study provide numerous directions for future research on the role of personal factors or candidate' beliefs,

characteristics and perspectives in the shaping of self-efficacy beliefs during the preparation phase. For example, although this study highlights that secondary teacher candidates might derive different experiences than their elementary or special education peers, it is beyond the scope of the study to explore the exact mechanisms that explain these differences. Future research might use mixed-methods to understand how teacher candidates in different teaching specialties perceive their teacher preparation experiences. Another direction for future research is to examine whether targeted supports can influence the relationship between personality traits and self-efficacy beliefs. For example, it appears from my findings that pre-service teachers who reported high levels of neuroticism at the beginning of the preparation program also reported the most increase in self-efficacy beliefs by the end of the program. It is possible that providing pre-service teachers who report higher-than-average neuroticism levels at the beginning of the program with additional student teaching experiences (or other forms of mastery experiences) may help reduce their anxiety and self-consciousness and develop their confidence in teaching.

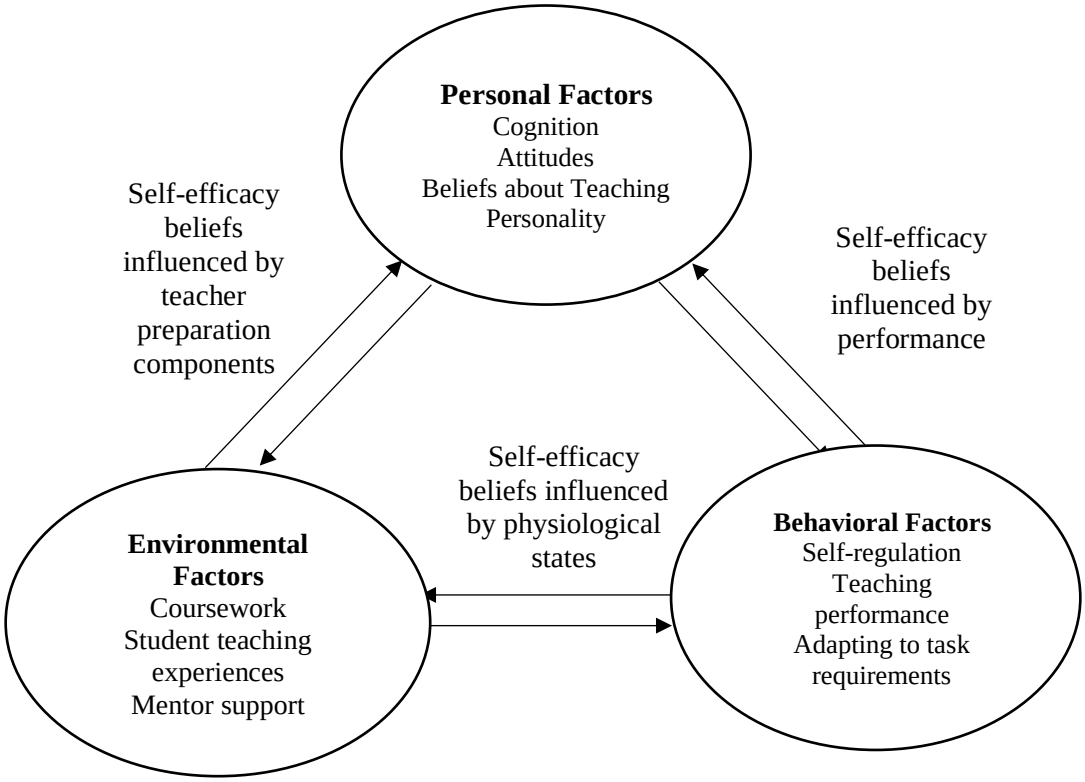
Conclusion

Teacher self-efficacy in general, and novice teacher self-efficacy in particular, is important for a variety of teacher practices and student outcomes (Tschannen-Moran & Hoy, 2001). Given that novice teachers' self-efficacy beliefs are largely formed during their training, it is important to pay attention to the role of personal factors in the development of self-efficacy beliefs in this stage. Learning more about the processes which underline self-efficacy might provide evidence on how to better support teacher candidates, increase their motivation and ultimately enable them to effectively teach in

classrooms (Jamil et al., 2012). Teacher education programs enroll a wide variety of applicants who bring with them different characteristics, personalities and ideologies with the goal of training them to teach in stressful and demanding classroom environments. This study contributes to the literature on self-efficacy beliefs among pre-service teachers and highlights the importance of understanding differences in candidates based on what they bring with them to the preparation program and the role these differences play in affecting their self-efficacy beliefs. Ultimately, helping teachers learn more about themselves and their beliefs may contribute to them developing more skills and confidence in their own ability as educators.

Figure 2

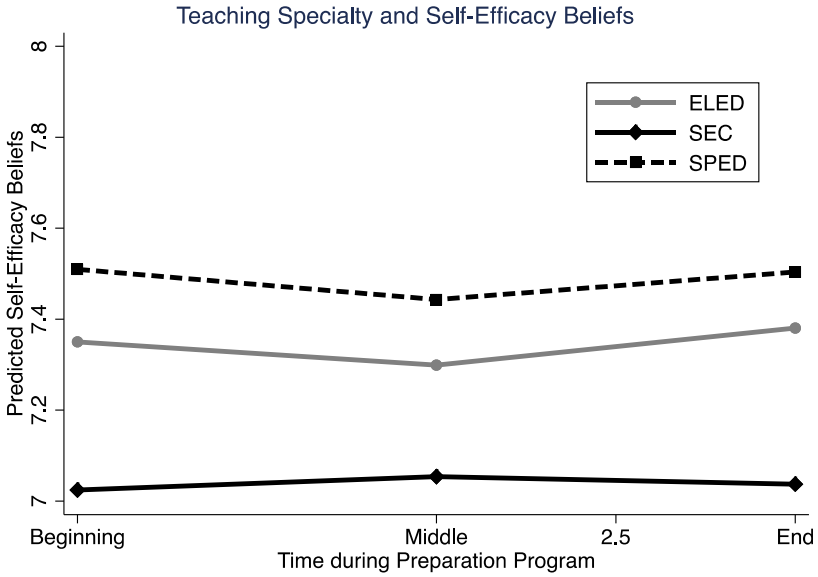
Social Cognitive Theory in the Context of Teacher Preparation



Note. Adapted from Bandura's Social Cognitive Theory

Figure 3

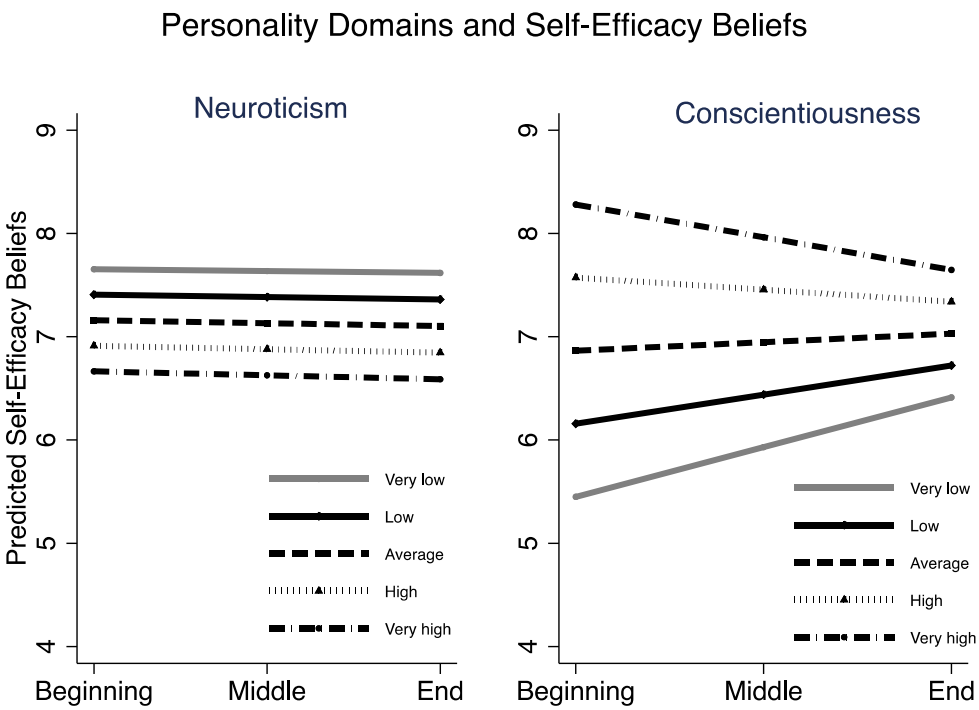
Effects of Teaching Specialty on Self-Efficacy Beliefs



Note. Model includes controls for demographic variables, personality domains, attitudes towards children, cohort and year fixed effects

Figure 4

Effects of Personality Domains on Self-Efficacy Beliefs



Note. Model includes controls for demographic variables, teaching specialty, other personality domains, attitudes towards children, cohort and year fixed effects

Table 1

Descriptive Statistics for Demographic Variables at Beginning of Preparation Program (N=515)

Demographic Covariates		**% Non-missing**	**%**
Program specialization		100%	
Elementary			41%
Secondary			46%
Special Education			13%
Gender		68%	
Female			55%
Male			13%
Race/Ethnicity		100%	
Caucasian/White			80%
Asian			10%
African American			5%
Hispanic			2%
Other			3%
Age		88%	
20 or under	45%		45%
21-23	41%		41%
24-26	1%		1%
27-29	0%		0%
30 or over	1%		1%
High School Location		100%	
Rural			19%
Suburban			71%
Urban			10%
High School Majority SES		100%	
Low SES			10%
Middle SES			64%
High SES			26%
High School Majority Race	100%		
Primarily students of color			7%
Mixed			42%
Primarily white students			51%

Table 2

Descriptive Statistics for Measures at Beginning of Preparation Program (N=515)

Measure	Mean (SD)	Range	Sample Min	Sample Max
1. Overall Self-efficacy	7.22 (0.93)	1-9	4.38	5
2. Neuroticism	2.63 (0.39)	1-5	1.36	4.09
3. Extraversion	3.34 (0.30)	1-5	2.58	4.33
4. Openness	3.01 (0.28)	1-5	2.08	4.33
5. Agreeableness	2.91 (0.32)	1-5	2.16	4.25
6. Conscientiousness	3.55 (0.24)	1-5	2.83	4.17
7. Ideas about Children	2.65 (0.32)	1-5	1.80	3.80

Table 3

Mean Self-Efficacy Levels at Time 1, Time 2 and Time 3 Of Preparation Program (N=1522 Across All Three Time Points)

	(1) Beginning of program (Time 1)	(2) Middle of program (Time 2)	(3) End of program (Time 3)
Overall Self-Efficacy	7.22 (0.93)	7.22 (0.92)	7.25 (0.95)

Note: Standard deviations are reported in parentheses

Table 4

Estimates of Fixed effects, Random effects and Goodness of Fit for Linear Growth Models (N=1522)

	Unconditional Model	Demographic variables Model 1	Adding Teaching Specialty Model 2	Adding Personality Domains Model 3	Adding Child-centered views of teaching Model 4
Fixed effects					
Intercept	7.23** (0.05)	7.11** (0.16)	7.30** (0.17)	5.24** (1.06)	3.70** (1.20)
Linear change	0.01 (0.02)	0.00 (0.02)	-0.02 (0.04)	0.43 (0.47)	0.42 (0.53)
Elementary			-	-	-
Elementary × Time			-	-	-
Secondary			-0.30* (0.12)	-0.27* (0.11)	-0.30** (0.11)
Secondary × Time			0.01 (0.05)	0.01 (0.05)	0.02 (0.05)
Special Education			-0.01 (0.16)	0.00 (0.16)	-0.01 (0.16)
Special Education × Time			0.08 (0.07)	0.08 (0.07)	0.08 (0.07)
Personality Domains					
Neuroticism				-0.26+ (0.14)	-0.24+ (0.14)
Neuroticism × Time				-0.02 (0.06)	-0.01 (0.06)
Extraversion				0.26 (0.18)	0.25 (0.18)
Extraversion × Time				-0.05 (0.08)	-0.05 (0.08)
Openness				-0.18 (0.19)	-0.09 (0.20)
Openness × Time				0.13 (0.09)	0.13 (0.09)
Agreeableness				-0.30+ (0.18)	-0.21 (0.18)
Agreeableness × Time				0.03 (0.08)	0.01 (0.08)
Conscientiousness				0.92** (0.23)	0.91** (0.23)
Conscientiousness × Time				-0.20* (0.10)	-0.20* (0.10)
Child-centered views of teaching					0.46** (0.17)
Child-centered views × Time					0.00 (0.08)
Level 2 (Random effects					
Intercept	0.81** (0.07)	0.75** (0.07)	0.74** (0.07)	0.63** (0.08)	0.61** (0.08)
Linear change	0.23** (0.04)	0.22** (0.05)	0.22** (0.05)	0.20** (0.05)	0.20**(0.05)

Rate of change	-0.46** (0.12)	-0.41** (0.15)	-0.42** (0.15)	-0.36 (0.21)	-0.36 (0.22)
Goodness of fit	365.34** (3)	323.71** (3)	310.96** (3)	239.65** (3)	226.41** (3)

** p<0.01 *p<0.05 +p<0.10

CHAPTER III: EXPERIMENTAL EVIDENCE ON THE ROBUSTNESS OF COACHING SUPPORTS IN TEACHER EDUCATION

In light of growing evidence that most novice teachers are forced to learn to teach "on the job", the field of teacher preparation has focused on how to push some of this learning into the pre-service period to provide teacher candidates with practical skills that they can apply in classroom contexts (Atteberry et al., 2015; Harris & Sass, 2011; Kraft & Papay, 2014; Papay & Laski, 2018). The need for robust preparation experiences is highlighted given that a majority of novice teachers report feeling underprepared and experience a reality-shock when they first enter the classroom environment, leading to stress, burnout, attrition and negative outcomes for students in the long term (Ingersoll, 2001; Papay & Laski, 2018). Optimizing preparation experiences to ensure that pre-service teacher candidates are ready to teach from their first day in the classroom has been a longstanding goal of preparation programs, yet there is a lack of evidence on the best ways to prepare new teachers in a robust, scalable manner.

All teacher educators agree that pre-service and novice teachers benefit from engaging in or doing the work of teaching, with feedback and support from teacher educators during the preparation experience (Ball & Forzani, 2009; Grossman, Hammerness & McDonald, 2009b; Hoffman, et al., 2015), As a result, university-based teacher preparation has shifted to a practice-based model with the ultimate aim of improved instruction and better student and teacher outcomes (Ball & Forzani, 2009). Despite the focus on situated practice with feedback in the preparation context, clinical

experiences and mentors vary in the amount of autonomy and practice opportunities they afford candidates. Not every candidate gets to practice every teaching skill they may need to engage and support students during their field placements (Ronfeldt, 2015). Thus, preparation programs across the country have been studying the degree to which some practice with targeted feedback can also occur in coursework. Such "approximations of teaching" -role-plays, rehearsals, and simulations-have been shown in qualitative work to support teacher candidates' ability to translate theoretical knowledge about "effective teaching" into their actual practice (Grossman et al., 2009b; Kavanagh & Rainey, 2017; Reisman et al., 2019; Windschitl, 2002).

One promising approach to providing candidates with feedback that is tied to practice opportunities during coursework (Kraft et al., 2018) is coaching. Coaching has been used extensively to support in-service teachers (Stahl et al., 2016) and has been shown to improve teachers' lesson planning and classroom organizational skills, use of instructional skills, as well as teachers' attitudes toward teaching, feelings of self-efficacy, and student achievement (Desimone & Pak, 2016; Kretlow & Bartholomew, 2010). Coaching can be a particularly powerful tool during the preparation phase since teacher candidates often lack the knowledge to identify strengths and weaknesses of their own (or their peers) while they are still learning to teach (Ericsson & Pool, 2016). Although the literature on coaching in the pre-service context is nascent, a handful of studies associate coaching with improvements in teacher candidates' affect, overall

satisfaction with preparation, attitudes towards self-development, and skill development (Bowman & McCormick, 2000; Cohen et al., 2020; Cohen & Wiseman, 2021).

What is less clear from the literature are the mechanisms, contexts and conditions that make coaching tied to practice opportunities effective during the preparation phase. Theories suggest that coaches serve as experts who can observe teachers as they carry out the work of teaching, evaluate their strengths and weaknesses and develop individualized strategies to skill development (Blazar & Kraft, 2015; Coburn & Woulfin, 2012), but there is hardly any empirical evidence on the efficacy of coaching supports for pre-service teachers (Kraft et al., 2018). The field of teacher coaching in general, and pre-service teacher coaching in particular, is in urgent need of rigorous evidence on the causal impact of providing pre-service candidates with coaching in the context of standardized practice opportunities on their pedagogical skills. In addition, researchers should explore the robustness of such coaching supports across systematic sources of variation we might expect to find given documented differences across preparation experiences and populations of teacher candidates (Blazar & Kraft, 2015; Hill et al., 2013; Ronfeldt, 2015) if we are to build the evidence-base on using supports such as coaching to prepare teachers for classroom teaching.

This paper examines the impact of a specific type of coaching provided by expert coaches, *directive coaching*, in improving pre-service teacher candidates' instructional practices during standardized, simulated practice opportunities. Along with providing evidence on whether directive coaching supports tied to practice opportunities are useful during the preparation phase, we also examine the robustness of coaching supports across

different teaching tasks, timing of study period, training context and modes of delivery within a single university site (Hammond & Moore, 2018; Ippolito, 2010). Using prospective replication research designs, we implemented a series of five randomized control trials (RCTs) at a single university site to evaluate the effectiveness of coaching in improving teacher candidates' practices in a mixed reality simulation setting. The simulations are considered "mixed reality" because the interface is a virtual classroom, but student avatars are remotely controlled by a trained actor, termed an 'interactor' (Dieker et al., 2013).

Across the five studies, we kept the coaching protocol and simulation setting consistent but systematically varied the teaching task being practiced in the simulator, the timing of when the study was conducted, and the mode of delivery of the simulated session. In addition, we tested whether coaching effects persisted for a group of undergraduate students who expressed an interest in teaching but were not yet enrolled in formal teacher preparation. We address two research questions: 1) What is the impact of targeted, directive coaching on pre-service teachers' skills in a simulated teaching environment? and 2) Are coaching effects robust across different sources of systematic variation (including different pedagogical teaching tasks, timing of the study period, context of practice and coaching and mode of delivery of simulated sessions)?

Overall, we find that providing pre-service teacher candidates with targeted, directive coaching significantly improves their pedagogical skills in a mixed-reality simulated practice setting. We also find that in the context of a single university site, coaching effects are robust in terms of direction, magnitude, and statistical significance patterns across different teaching tasks, timing of the coaching intervention, and different

modes of delivery. However, we find that coaching effects are smaller in magnitude (yet significant) for participants not yet enrolled in teacher preparation. Findings from this paper have significant implications for teacher preparation programs and educators as the field explores scalable innovations for enhancing the preparedness of new teachers.

This paper proceeds as follows. In Section 1, we provide background and an overview of the literature on practice-based teacher education, and coaching supports to improve teaching practices. Here, we also describe sources of variation that we use to examine patterns of robustness in coaching effects. In Section 2, we outline our research methods, including sample and setting characteristics, treatment and control contrast, experimental design, conceptual replication designs and our analytic plan. In Section 3, we present our results estimating the impact of providing participants with targeted, directive coaching on their teaching practices in a simulated teaching context and assess the robustness of coaching effects across systematic sources of variation. In Section 4, we discuss our findings, provide implications for teacher preparation policy and directions for future research.

Background

The literature has consistently documented that novice teachers struggle during the initial period of teaching (Friedman & Gavish, 2001), often using terms such as "reality shock" (Veenman, 1984; McCormack & Thomas, 2003) or "transition shock" (Corcoran, 1981) to describe the mismatch between novice teachers' expectations and the reality they encounter when they first enter a classroom. In part, this mismatch is due to tremendous variation in how pre-service teachers are prepared, leading to successive cohorts of new teachers entering the classroom with vastly different knowledge and skills

(Ronfeldt, 2015). Criticisms against teacher preparation programs include a narrow focus on helping teacher candidates learn "what to teach", rather than addressing what being a teacher really means and how to transfer generalize acquired teaching techniques to real world settings (situated practice, importance of simulated sessions mirroring what happens in real classrooms) (Scheeler, 2007; Gavish & Friedman, 2010). Evidence suggests that most novice teachers improve dramatically during their first few years, learning a lot on the job (Atteberry et al., 2015; Harris & Sass, 2011; Kraft & Papay, 2014; Papay & Laski, 2018). From the teacher preparation perspective, this raises the question of whether (and how) we can push some of this learning into the pre-service period so that new teachers are better prepared from their first day in the profession.

Practice opportunities and feedback in teacher preparation

Traditionally, teacher preparation programs across the country have employed an "apprenticeship" model where teacher candidates learn to teach by observing teacher educators in clinical settings and other experts do the work of teaching while simultaneously developing content and pedagogical knowledge through their coursework (Cruikshank & Armaline, 1986). While useful in giving teacher candidates a sense of instructional techniques, apprenticeship models are often problematic because they sometimes reinforce incorrect instructional practices and even counter the teachings provided in coursework (Grossman et al., 2009b). In contrast, practice opportunities grounded in university classrooms can provide candidates flexibility to experiment, learn, regroup and reflect on their own practice rather than shadow performance by mentor teachers or other experts in clinical settings.

Recently, teacher preparation programs have shifted to practice-based approaches in teacher education in an effort to provide pre-service teacher candidates with more opportunities to practice their instructional skills and get a feel for what being a lead teacher in a classroom environment involves, while also receiving feedback from their mentors (Levine, 2006). A practice-based approach (PBTE) to teacher preparation centers on the idea that candidates need scaffolded experiences to develop classroom-skills, in addition to more traditional approaches that foreground educational theory and subject-area knowledge (Grossman et al., 2009b). With the move to PBTE, teacher educators have begun experimenting with providing candidates with "representations" or examples of skilled teaching during their coursework that are then "decomposed" or broken apart to highlight key features of the particular teaching practice (Grossman et al., 2009b).

Research on practice-based teacher education highlights the importance of two critical "components" that allow teacher candidates to improve their instructional practices during the preparation phase- practice and targeted feedback. Specifically, teacher candidates benefit from deliberately sequenced practice opportunities that increase in complexity over time, along with extensive feedback from an expert teacher or mentor (Carpenter & DeLosh, 2005; Karpicke & Bauernschmidt, 2011) when learning complex skills required for effective teaching in a classroom setting. What remains unclear is the context in which such practice opportunities and targeted feedback should occur and the best ways to couple practice opportunities and feedback supports to improve teacher candidates' practice (Lampert, 2009). Given the limited duration but critical nature of the teacher preparatory period, the field is need of more rigorous and generalizable evidence about the best ways to support teacher candidates through

standardized practice sessions and feedback to improve their instructional skills during the pre-service phase.

Coaching to Improve Instructional Skills

One way to provide teacher candidates with a combination of practice opportunities and targeted feedback for improvement is through coaching. Typical models of coaching for pre-service teachers incorporate feedback from three sources: the university supervisor, the directing teacher or a peer coach (Anderson & Radencich, 2001). In such contexts, mentor teachers, university supervisors and instructors work together to provide support and feedback (most often in group settings) to improve teacher candidates' success during their field placements (Darling-Hammond, Holtzman, Gatlin & Heilig, 2005; Scheeler, 2007). Unfortunately, this feedback is often delayed and occurs infrequently because of the time and resources that supervisors and mentors need to devote to the process, which makes coaching from supervisors inadequate for skill development among pre-service teacher candidates. Peer coaching, where pre-service teacher candidates observe each other practicing instructional skills and offer feedback in group settings, is another form of preservice teacher coaching that has been shown to improve skills in the short term (Britton & Anderson, 2010), but is not as effective at improving candidates' instructional skills as receiving feedback from experts or mentors with more experience and knowledge (Hoffman et al., 2015).

Coaching is a popular strategy for improving in-service teacher practice (Coburn & Woulfin, 2012), where instructional experts work with teachers to improve specific aspects of their teaching, either individually or as part of a small group (Fletcher & Mullen, 2012). The literature on coaching has broadly distinguished between two types of

coaching based on how coaches position themselves in relation to teachers- responsive coaching (Dozier, 2006) and directive coaching (Deussen, Coskie, Robinson & Autio, 2007). Responsive coaching relationships are where coaches facilitate self-reflection among teachers, thereby allowing teachers' needs to guide the process (Ippolito, 2010). In contrast, directive coaching occurs when coaches play the role of an expert, identify teachers' areas of weakness and provide specific strategies for improvement (Deussen et al, 2007; Hammond & Moore, 2018; Ippolito, 2010) related to specific instructional skills. Previous work suggests that directive coaching works better with less experienced teachers (Killion, 2016; Steiner & Kowal, 2007), who benefit from specific, targeted and individualized feedback on their instructional performance or practice.

Despite theoretical knowledge that pre-service teachers benefit from immediate, personalized feedback from "expert others" to improve their instructional practice, the empirical evidence on how to structure and implement coaching supports in the context of standardized, sequential practice opportunities during teacher preparation remains weak. This lack of evidence around coaching in the pre-service context is surprising, given that when learning complex teaching skills, teacher candidates would likely benefit even more from learning opportunities that involve extensive scaffolding that coaching provides (Karpicke & Bauernschmidt, 2011) than in-service teachers. During the preparation phase, candidates often do not have sufficient knowledge to recognize their own strengths and weaknesses related to teaching, to isolate strategies that result ineffective instruction or to fully understand the impact of their instruction on students (Ericcson & Pool, 2016). Here, a coach who provides specific, targeted feedback that focuses on a narrow set of instructional skills (as in the case of directive coaching) may

have the power to transform how pre-service teachers learn, and their instructional efficacy when they enter the classroom.

The potential of simulated teaching environments

Given the limited duration of teacher education, policy makers and programs alike need more empirical evidence to make strategic choices about the nature, structure, and sequence of practice opportunities that can incorporate feedback in the form of coaching. Current practice opportunities during the pre-service phase in the form of "role-playing" as part of group activities or during student teaching placements are not realistic, feasible or scalable in a standardized manner across all teacher candidates (Ronfeldt, 2015). Digitally mediated simulations that have been used widely in other professions such as aviation and the military offer realistic and standardized spaces to approximate practice that can be embedded into coursework that pre-service teacher candidates routinely complete as part of their preparation experience.

In teacher preparation, mixed reality simulations can be designed to reflect the real-world complexities of live teaching environments (Slater, 2009), especially since studies have shown that simulations feel more realistic than other approximations of practice including role-plays and rehearsals, and that participants respond in ways that are closely aligned with what they might do in actual classroom settings (Arora et al., 2011a; Dieker et al., 2013; Slater, 2009). Voice actors (termed "interactors) who control digitally mediated avatars can be trained to respond in real time to teacher candidates' instructional cues in ways that real children would. Importantly, teacher candidates are able to suspend disbelief and have a complete practice experience since they do not know

exactly how the technology works, or that a person is controlling the simulated student avatars.

Technologically mediated simulations are particularly promising in the context of causal work, since they afford standardized opportunities across teacher candidates in a cost-and-time effective manner (In this study, participants completed simulated sessions that were each on average, 15 minutes long, including targeted coaching supports). One of the major reasons for the lack of causal evidence surrounding the efficacy of practice opportunities with coaching during the pre-service phase is the inability to control for differences in clinical placements (Bell, Jones, Qi & Lewis, 2018), as well as logistical and methodological challenges with observing pre-service teacher candidates in live classroom settings. By allowing teacher educators and researchers to deliver simulated practice settings in exactly the same manner for all candidates, there is clearer focus on specific instructional skills that is not confounded by variation in teacher candidates themselves, preparation experiences or mentor teachers (Steinberg & Garrett, 2016). Finally, the short duration of simulation sessions allow teacher candidates the opportunity to "do-over" their practice, along with targeted supports such as coaching in an efficient and scalable manner.

Efficacy of coaching supports during simulated practice opportunities

In previous work, we examined whether pre-service teacher candidates in two independent studies showed improved skills when they received targeted, directive coaching from trained coaches in a simulated classroom (Cohen et al., 2020; Cohen & Wiseman, 2021). In one study where teacher candidates focused on redirecting off-task student avatar behaviors, we randomized 105 candidates who completed simulator

session as part of their methods coursework to one of three conditions: bug-in-ear feedback with coaching, coaching and self-reflection (Cohen et al., 2020). In our findings, we observe that both bug-in-ear feedback with coaching and coaching only produced statistically significant effects on teacher candidates' performance at redirecting off-task behaviors in the simulator. The bug-in-ear feedback with coaching group scored 2.50 points higher on a 10-point quality performance scale relative to the self-reflection group (ES = 1.86 SD; $p < 0.01$), while the coaching group scored 2.54 points higher than the "Self-Reflection" group (ES =1.70 SD; $p < 0.01$). In another study, 103 candidates were (independently) randomized to either complete self-reflection or receive targeted feedback from an expert coach as they practiced providing high-quality feedback to simulated student avatars during a text-based discussion (Cohen & Wiseman, 2021). Results suggest that candidates who receive coaching performed significantly better than their peers in the self-reflection group. Specifically, coached candidates scored 1.30 points higher (ES=1.25 SD, $p<0.01$) on an overall quality of feedback scale, compared to their peers who instead engaged in self-reflection. Taken together, findings from both studies provide encouraging evidence that providing pre-service teacher candidates with targeted, directive coaching can improve their proximal outcomes of teaching skills and beliefs.

However, both studies are limited because they each examined the efficacy of coaching for a specific teaching practice (either setting classroom norms or providing high-quality feedback) for a single group of teacher candidates. Thus, although we found that coaching can help improve skills with which novice and pre-service teachers often report struggling, there is need for more evidence on whether these effects replicate

across different systematic sources of variation (such as different instructional practices, settings and modes of delivery). Building the empirical evidence base on the best ways to provide teacher candidates with simulated practice opportunities coupled with directive coaching is critical for generalizing effects in teacher preparation (Cole & Stuart, 2010; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton & Olson, 2018).

In this paper, we extend our work to examine the *contexts and conditions under which* directive coaching improves pre-service teacher candidates' skills during simulated practice opportunities. One way of examining why or what might make coaching more or less effective for certain pre-service teachers is to systematically introduce different sources of variation that might explain heterogeneity in coaching effects, while keeping the coaching protocol consistent. In this paper, we analyze data from five[3] studies that evaluate the impact of providing participants with expert coaching on improving their pedagogical skills in the simulator. Across the five studies, we are interested in whether coaching effects are robust across four specific sources of variation: 1) different teaching tasks, 2) different study timings, 3) different contexts for simulated practice and directive coaching supports and 4) different modes of delivery of simulated sessions and coaching.

Teaching task. We first examine differences in coaching effects across different pedagogical tasks practiced in the simulator. A majority of the research on coaching in-service teachers focuses on content-specific coaching programs in reading, literary, science or math. In the meta-analysis by Kraft and colleagues (2018) on the impact of coaching supports for in-service teachers, the authors find smaller coaching effects on teachers' instructional practices and student achievement for coaching programs that

[3] The research team also conducted an additional replication RCT in Fall 2019. We do not include data from that study in this paper because of a change in the coaching protocol.

focus on general instructional pedagogies (0.07 SD) than content-specific programs (0.20 SD). This narrow focus on coaching for content-specific practices precludes knowledge of whether different instructional practices are affected differently by coaching for the same individuals. Depending on the underlying theory of action, the same type of coaching may have differential impacts on different areas of teachers' practice, such as on content-specific tasks and general tasks related to classroom management. Typically, in order to improve instructional practices in specific content areas such as reading or literacy, the literature suggests that teachers as well as coaches need deep subject or content area knowledge in addition to general teaching methods to benefit most from coach feedback (Vogt & Rogalla, 2009). In contrast, coach feedback focused on general pedagogical practice does not focus as much on subject area, but on teacher candidates' knowledge of general classroom instruction and management.

Timing of study. We then look at differences in coaching effects across two consecutive cohorts of pre-service teachers at the same university site. Although there are a few empirical studies on the effectiveness of coaching across multiple cohorts (see Blazar & Kraft, 2015; Killion, 2016), the focus of these evaluations has largely been to assess the impact of changes in the coaching model used across cohorts rather than to assess the robustness of the same coaching model on different samples of participants. To build the empirical evidence base, it is important to systematically investigate whether coaching supports in the context of simulated practice opportunities consistently improves instructional skills among participants who are similar in all aspects other than the timing of the study.

Training context. One of the central issues for teacher education is how to foster learning in the context of practice (Helgevold, Næsheim-Bjørkvik & Østrem , 2015) and Darling-Hammond et al. (2005) argue that clinical experiences should be carefully mentored. For coaching to be effective during pre-service teacher education, the literature suggests that coaching (as well as practice opportunities more broadly) should be accompanied by instruction that focuses on evidence-based practices, as well as the opportunity to concurrently implement these supports either in a practicum or student teaching setting (Elmore, 2006; Scheeler, Bruno, Grubb & Seavey, 2009). That is, approximations of practice, such as simulations, should not be stand-alone experiences. In their meta-analysis, Kraft et al. (2018) find that in-service coaching paired with some additional training (either in the form of group training workshops or coursework in the pre-service context) is associated with a stronger coaching effect (0.31 SD larger for instructional practices and 0.12 SD larger for student achievement). Because a majority of coaching programs are usually accompanied by some form of training (Kraft et al., 2018), it is difficult to understand the efficacy of the same coaching program in the absence of additional training supports.

Mode of delivery of simulated sessions. Finally, we compare coaching effects across in-person and virtual simulated sessions to understand whether coaching is more or less effective when delivered virtually. Virtual or online coaching programs could provide a more resource-effective way of providing both in-service and pre-service teachers with feedback to improve their skills (see Israel, Knowlton, Griswold & Rowland, 2009; Rock et al., 2013; Stapleton, Tschida & Cuthrell, 2017) since it allows for a fewer number of coaches to provide feedback to a larger number of teachers without

geographical constraints. To be successful, Israel and colleagues (2009) suggest that virtual or online coaching must mimic face-to-face coaching in that it must be immediate and tied to a real educational context (Rock et al., 2013). In the Kraft et al. (2018) meta-analysis, the authors examine coaching effects from 47 coaching programs that provide teachers with face-to-face coaching and 13 programs that use virtual or online coaching. They do not find statistically significant differences in effect sizes between the two modes of delivery, but acknowledge that they are likely underpowered to detect meaningful differences.

Through a series of experimental studies in this paper, our goal is to build evidenced-based theory on how teacher preparation programs and teacher educators can use simulated practice and directive coaching to improve pre-service teacher candidates' instructional skills. We begin by examining the impact of providing participants with simulated practice sessions and directive coaching across five RCTs that were designed to be prospective replication studies. We then extend our findings to assess whether coaching effects are persistent across different sources of variation including different teaching tasks, different timing of the study, different contexts for the practice sessions and different modes of delivery.

Research Methods

To examine the efficacy and robustness of coaching supports during the teacher preparation phase, we use data from five[1] individual randomized control trials (RCTs) to construct systematic conceptual replication studies. All five studies were implemented in the context of a single university-based teacher preparation program, and focused on evaluating the impact of providing pre-service teachers with targeted coaching supports

to improve their pedagogical practice in a simulated classroom setting. As outlined in Figure 5, the five[1] RCTs were conducted during the fall and spring semesters across three academic years (2017-2018 to 2019-2020). In this section, we provide an overview of the samples and settings, experimental design as well as an overview of measures for the experimental studies included in this paper. We also outline the treatment (coaching) and control contrast and provide an overview of how we constructed replication studies using data from the individual experimental studies. Finally, we lay out our analytic plan for assessing the efficacy and robustness of directive coaching supports on pre-service teacher candidates' instructional skills.

Samples and Settings

All five experimental studies were conducted at a large, selective, public university in southeast United States. Participants in four of the five experimental studies (studies 1, 2, 3 and 5) were enrolled in the university's teacher preparation program that prepares approximately 100 new teachers each year. Participants in the remaining study (Study 4) were enrolled in the same university, but were not in the teacher preparation program. They were recruited through an undergraduate course for students interested in exploring teaching as a profession. Across the different studies, the teacher candidate samples were largely White (81%), female (88%) and middle class (61%). The undergraduate sample was less White (56%), almost equally female and male (50%) and predominantly middle class (57%). Approximately 43% of the undergraduate sample reported an interest in teaching as a possible profession and about 63% reported some prior experience with children (either as a babysitter, coach, mentor or camp counselor). Mean characteristics of participants in each of the five studies are presented in Table 5.

Experimental Design and Data Collection Procedures for Individual Studies

Figure 6 summarizes the experimental design and study protocol for each of the five studies. For studies 1, 2, 3, and 5, pre-service teacher candidates engaged in a series of simulation sessions. The simulations were designed to be parallel in that at different points during each study, participants were exposed to different simulated student avatar responses (within each teaching task) while keeping the number and type of responses standardized across sessions. As an example, participants who practiced the classroom norms task were exposed to different "off-task" behaviors in each simulation session, but the number, type and severity of the behaviors were kept consistent across sessions.

For each study, candidates completed a baseline simulation session (Time 1) during which the research team collected covariate information about their baseline characteristics and pretest measures of the outcomes. Candidates were then randomly assigned within their course sections to one of two intervention conditions: coaching or self-reflection between simulation sessions. Although coaches and simulation specialists were not randomly assigned, they were scheduled in a manner that ensured sufficient variation across course sections, days and timings of sessions, allowing the research team to control for potential coach and interactor effects.

Approximately two months after their baseline sessions, participants completed a second simulation session (at Time 2) where their teaching outcomes were observed and recorded. Immediately after the simulation session at Time 2, participants randomly assigned to the coaching condition received five minutes of face-to-face coaching with an expert coach, while participants assigned to the self-reflection condition completed a series of reflection prompts designed by the research team without any support from a

coach. After the five-minute coaching or self-reflection block, participants completed a third parallel simulation session without any support. Candidates' performance during the simulator sessions were observed and recorded at all three time points.

The protocol for Study 4 deviated slightly from the other four studies. This study did not take place in context of a teacher preparation program but in an undergraduate class for students interested in exploring the teaching profession. In this study, participants completed simulation sessions at times 2 and 3, but they did not complete a baseline simulation session. Covariate information about these participants' baseline characteristics were collected as soon as participants consented to participate in the study, using the same surveys that were administered to teacher candidates (to allow for consistency in measures). After collecting this information, undergraduate participants were randomized within their course sections to self-reflection or coaching prior to their first simulation session at Time 2.

Treatment Contrast

As part of their simulator sessions, teacher candidates were randomly assigned to either complete self-reflection on their performance in the simulator, or to receive targeted, individualized, directive from an expert coach. A team of trained expert coaches observed a participant's performance in the simulator and provided them with five minutes of feedback regarding their performance. Each coaching conversation followed a similar coaching structure: coaches would first gauge participants' understanding of their own performance, identify elements of effective instruction using a skill progression designed by the research team for each scenario, highlight a specific skill in need for improvement, and finally reinforce the link between the skill and specific instructional

strategies using a teacher-coach role play (Desimone & Pak, 2016). Each participant then had another opportunity to practice their skills in the simulator immediately following their conversation with the coach.

Coaching condition. Across all studies, coaches observed participants' performance in the simulator at Time 2 and Time 3. After the simulated session at Time 2, the coach provided participants who were randomly assigned to the coaching condition with five minutes of targeted, directive feedback about their performance. All coaches were doctoral students in education who had trained intensively with the research team to ensure that the structure of the coaching conversation was consistent across all participants. First, the coaches gauged participants' understanding of their performance and then they identified and affirmed positive elements of the participant's practice; third, they supported the participant in developing a specific, targeted skill (identified using a skill progression for each pedagogical task designed by the research team) and finished by role-playing with the participant, thereby allowing them to practice and rehearse the skill before they completed the simulated session at Time 3.

Self-reflection (business-as-usual) condition. Participants who were assigned to self-reflection were asked to reflect on their performance in the simulator using a self-reflection protocol designed by the research team. After participants completed the simulated session at Time 2, they completed the self-reflection protocol before they practiced their skills in the simulator once more at Time 3. The self-reflection protocol was designed using literature on teacher self-reflection (Yost, 2006) and included three questions: "What are some things you think went well in terms of redirecting student behavior?" "What are some things you think could have gone differently in terms of

redirecting student behavior?" and "What are you going to work on in the next 5-min session to improve your redirections of student behavior?"

Measures

We used a multi-method approach to data collection across the five RCTs, which allows a more valid and reliable assessment of constructs that we are interested in. We chose the following measures because of their strong reliability and validity evidence for use with pre-service teachers (as summarized in Table 6), their ease of use and training, and their demonstrated sensitivity to the coaching intervention (see Cohen et al., 2020).

Candidate characteristics. Demographic and baseline information about teacher candidates in studies 1, 2 3 and 5 were collected from the teacher preparation program's longitudinal data system. To keep the measures identical across studies, participants in the undergraduate teaching as a profession course as part of Study 4 were administered the same surveys at the beginning of the study period. Surveys included baseline information about participants' undergraduate major and high school GPA, parental education and characteristics of the high school attended (average achievement level, average SES level and urbanicity of school). Participants also completed measures of personality including the NEO Five Factor Inventory (McCrae & Costa, 2004), Teacher Sense of Self-Efficacy (Tschannen-Moran & Hoy, 2001) and multi-cultural attitudes (Munroe & Pearson, 2006). A descriptive summary of measures and their psychometric properties is provided in Table 6.

Baseline simulation performance information. At the start of each study period, teacher candidates (in studies 1, 2, 3 and 5) completed pre-intervention simulation sessions. These sessions occurred at Time 1 in Figure 6 and data from these sessions are

used as pretest measures of teacher candidate performance in the simulator described below. As mentioned earlier, undergraduate participants in Study 4 did not complete baseline simulation sessions.

Teacher candidate performance outcomes. Our outcome of interest is derived from observational rubrics of teacher candidates' simulator performance. The research team designed an observation protocol to examine the impact of coaching on participants' instructional and pedagogical skills based on the Responsive Classroom (2014) framework. A team of trained and certified raters who were blind to participants' randomized intervention condition and which simulation session they were coding, scored videos of candidates' simulated sessions from the three time points. Reliability using Krippendorff's alpha was calculated for a subset of double-coded videos and ranged from 0.75 to 0.88 across the different studies (Krippendorf, 1970). Coder drift was addressed using weekly calibration checks and agreement reports run by the research team, and scores from double-coded videos were averaged to form aggregate measures of performance. Although coders scored videos for a variety of different outcomes across the two simulation tasks, we use a measure of overall quality of performance as our primary outcome of interest in this paper. For the overall quality score, coders assigned each simulation a score (ranging from 1 to 10) that reflected the extent to which teacher candidates met the goals of the simulated scenario (either provided high-quality feedback or effectively redirected off-task behaviors).

Causal Replication Framework and Research Designs

To assess the contexts and conditions under which providing pre-service teacher candidates with targeted coaching supports improved their instructional performance

during the simulated teaching sessions, we use the Causal Replication Framework (CRF) introduced by Wong and Steiner (2018) to derive replication designs with the five individual experimental studies. The CRF allows researchers to identify boundary conditions for replicability of effects across units, treatments, outcomes and settings (UTOS) for the purpose of generalizing scientific findings (Nosek & Errington, 2020) and describes five assumptions required for treatment effects to directly replicate presented in Appendix Table 1- they require that all studies have the same treatment-control contrast, measure the same outcomes, and have similar distributions of unit and setting characteristics that may moderate effects; individual study assumptions require that effect estimates for each study are unbiased and correctly reported. By systematically varying study features including contexts, settings and participant characteristics in this paper, we are able to identify the source of coaching effect variation which is essential to understanding the best ways that directive coaching supports and simulated practice opportunities can be used to improve candidates' instructional skills during the preparation phase.

Using the CRF, we employed three types of replication designs (using the five individual RCT studies) that included a switching replication design, a multiple cohort design, and a "matched" conceptual replication design, which varied the target population and setting under which the coaching intervention was introduced (Wong, Steiner & Anglin, 2020). Each of the study designs and corresponding RCTs are described in detail below, and a summary of replication assumptions systematically tested by each research design is provided in Appendix Table 2.

Switching replication design. To examine the robustness of effects across different pedagogical tasks, we used a switching replication design (with data from Study 2 and Study 3) where teacher candidates were randomly assigned in the fall semester (Study 2) to receive either coaching or engage in self-reflection related to facilitating a text-based discussion in the simulated classroom environment. In the spring semester (Study 3), the *same group of candidates* practiced setting classroom norms and redirecting off-task behaviors in the simulator. Between the fall and spring semesters, candidates' coaching condition was "switched" relative to what they received in the fall semester. In practice, teacher candidates who received coaching in the fall engaged in self-reflection in the spring and vice versa. As summarized in Appendix Table 2, shared participants across studies 2 and 3 ensure the same distribution of teacher candidates, and study procedures including random assignment, sensitivity tests in the estimation models, and independent reporting of results ensure that each study result was reported without bias. Finally, the research team empirically assessed the extent to which outcome measures and intervention conditions were stable across studies.

By comparing coaching effects across studies 2 and 3, we assess the robustness of the impact of coaching across a subset of pedagogical tasks. Specifically, we examine coaching effect variation across the two different teaching tasks- 1) providing high-quality feedback in a text-based discussion and 2) redirecting off-task student behavior during a norm-setting exercise. Since assumptions related to units, treatments and outcomes were met across the two studies included in the switching replication design, we hypothesize that any differences in coaching effects from the two studies can be attributed to the difference in the pedagogical task being practiced in the simulator. If our

results suggest coaching effects are robust across the two tasks, we have evidence that coaching programs improve participants' teaching skills across at least some subset of pedagogical tasks in the context of simulated practice.

Multiple-cohort design. To examine the robustness of effects across time (and small insignificant changes in teacher candidate composition), we examined coaching effects across two cohorts of teacher candidates (Study 1 versus Study 3). Multiple cohort studies are useful in examining the effectiveness of treatments across different cohorts and samples of participants, where the treatment protocol is delivered consistently for all cohorts. By comparing coaching effects across multiple cohorts of participants at the same institution, we can assess whether the same coaching intervention is more or less equally effective over time. Here, we compare coaching effects for a cohort of teacher candidates in Study 1 with coaching effects for a second cohort of teacher candidates in Study 3 who completed simulated sessions a year later. In both studies, teacher candidates practiced setting classroom norms and redirecting off-task behaviors displayed by the student avatars, and received face-to-face coaching from trained coaches who followed the same coaching protocol.

Across both studies, cohorts of participants were enrolled in the same preparation program, received the same training, participated in standardized simulated sessions that were designed to be parallel to each other in structure and format (meaning that teacher candidates had the same number of opportunities to respond to student avatars across the two studies), focused on the same pedagogical task (setting classroom norms) and received face-to-face coaching from an expert coach who followed identical coaching protocols designed by the research team. If our results suggest coaching effects are robust

across cohorts, then we hypothesize that targeted, directive coaching may consistently improve participants' teaching skills across time when samples are similar on observable characteristics.

Matched Conceptual Replication Designs. We then used a matched replication design to assess the robustness of coaching effects with a different training context (Study 3 versus Study 4), as well as across different modes of delivery of the simulation session (Study 3 versus Study 5). The goal for these studies was to systematically vary study characteristics (including units and settings) under investigation while "matching" other study factors that were not under investigation (Wong & Steiner, 2018). For example, when comparing coaching effects across different training contexts, participants in Study 4 were not teacher candidates enrolled in the teacher preparation program (as in the case of Study 3), but undergraduates enrolled in an introductory teaching course exploring teaching as a profession. However, the participants in Study 4 engaged in the same simulation scenarios, received the same coaching protocol, and were administered the same outcomes as participants in Study 3. In studies 3 and 5, two cohorts of teacher candidates practiced the classroom norms scenario, but the second cohort of candidates completed simulation sessions via zoom and not in-person (as with Study 3). Across both sets of comparisons, the study team assessed "balance" in covariate characteristics across studies to examine the extent to which other replication and individual study assumptions were met (or violated). These are summarized in tables 5, 7, 8 and Appendix Table 2.

Robustness of coaching effects across different training contexts. We assess the robustness of coaching effects across training contexts by comparing coaching effects for a sample of pre-service teacher candidates enrolled in the teacher preparation program

with coaching effects (Study 3) for a sample of undergraduate students enrolled in an introduction to teaching course at the same school of education (Study 4). For the sample of teacher candidates, the simulated sessions were designed to complement material covered during general methods courses in which they were concurrently enrolled. Teacher candidates were therefore familiar with the content and pedagogical knowledge underlying the simulator scenarios. In contrast, the introduction to teaching course was designed to provide undergraduate students with a sense of teaching as a profession, but did not cover any specific teaching practices in detail. Because both samples of participants completed standardized simulated sessions that were designed to be parallel to each other in structure and format, focused on the same pedagogical task (setting classroom norms), completed sessions in-person and received the same coaching protocol, we hypothesize that any observed differences in coaching effects can be attributed to differences in sample characteristics (for example, demographic characteristics of the participants themselves), setting characteristics (such as the prior training each set of participants received on setting classroom norms) or a combination thereof.

To disentangle the influence of sample characteristics from training context on coaching effects, we used propensity scores to "reweight" our sample of undergraduate students such that they are more similar on observable characteristics to our sample of teacher candidates. A summary of the weighting procedures in given in Appendix A. The difference in coaching effects between the resulting weighted samples allows us to better isolate and identify whether coaching effects are robust across different training contexts, as long as participants who receive coaching are similar on observable characteristics.

Robustness of coaching effects across different settings. Next, we compare coaching effects across two studies (and two cohorts of pre-service teacher candidates) where one sample of pre-service teachers completed simulation sessions in person (Study 3) and received face-to-face coaching and the other sample of pre-service teachers completed simulation sessions over a virtual platform and received virtual coaching in real-time (Study 5). Both samples of pre-service teachers were enrolled in the same methods courses, practiced the same pedagogical task in the simulator, completed parallel simulated sessions (in that they responded to the same number and similar prompts from the simulated students), received comparable coaching from trained expert coaches and were similar on observable characteristics. Here, we hypothesize that any differences in coaching effects observed between the two samples are likely due to whether participants completed simulated sessions in-person or online.

Analysis

To examine the efficacy of directive coaching supports in pre-service teacher education and the robustness of coaching effects across different sources of variation, our analysis consisted of four steps. First, we conducted diagnostic checks of our research design assumptions for each experimental study as well as the replication assumptions laid out in the CRF (Wong & Steiner, 2018) to identify specific sources of variation across the replication studies. Next, we estimated the average treatment effect (ATE) of coaching on participants' performance in the simulator for each of the individual studies, which allows us to provide causal evidence on whether directive coaching improves participants' instructional performance in the simulator. We also estimated the pooled coaching effect across the five studies using meta-analytic methods and assessed whether

we had evidence of coaching effect heterogeneity across the five studies using the I^2-statistic and Q-test (Higgins, 2003). Finally, we compared coaching effects across different pairs of studies using distance-based correspondence measures (including magnitude of effects, direction of effects and statistical significance patterns) to identify patterns of robustness.

Baseline equivalence across studies. We began by examining baseline equivalence (data from Time 1 for our teacher candidate samples and Time 2 for the undergraduate sample) across treatment groups in each of the replication studies. Specifically, we examined differences in participant and setting characteristics across sites and looked at patterns of attrition, missing data and treatment non-compliance. These checks included factors that we systematically varied across studies in order to uncover sources of variation (which we refer to as systematic variation factors) and factors that we kept constant across studies (which we refer to systematic replication factors). Tables 5, 7 and 8 present the results of our balance checks. Appendix Table 2 summarizes the replication and individual study assumptions across the three research designs used in this paper.

Impact of coaching on participants' instructional practices. After examining baseline equivalence of systematic variation factors and systematic replication factors, we estimated the effect of coaching on participants' instructional practices at Time 3 for each study. Specifically, we estimated the average treatment effect (ATE) using the following model:

$$Y_{ij(t=3)} = \beta_0 + \beta_1 Coaching_{ij(t=3)} + (X_{ij})\beta + \alpha_j + \epsilon_{ij} \quad (7)$$

In Equation 7, $Y_{ij(t=3)}$ represents the overall quality score for participant *i* in course section *j* at Time=*3* and is a function of participant *i*'s treatment status (where *Coaching*=1 if assigned to receive coaching and *Coaching*=0 if assigned to participate in self-reflection), and individual characteristics (X_{ij}) measured at baseline. The model also controls for course section fixed effects (α_j). The coefficient β_1 represents the average treatment effect (ATE) of coaching for each study. Results from Equation 7 are presented in Table 9. Table 9a includes treatment effects for the propensity weighted samples.

Meta-analytic coaching effect across studies. Next, we estimated the cross-site pooled effect using meta-analytic methods that account for the clustered nature of the data and produced precision weighted estimates (Hedges, Tipton & Johnson, 2010). We used a fixed effects meta-analytic model where effect sizes represent the average treatment effect (ATE) of coaching on participants' instructional practices using the following model:

$$Y_{ijs(t=3)} = \beta_1 Coaching_{ijs(t=3)} + (X_{ijs})\beta + \alpha_s + \lambda_j + \epsilon_{ijs} \quad (8)$$

In Equation 8, $Y_{ijs(t=3)}$ represents the overall quality score for participant *i* in section *j* and study *s* at Time=*3*. $Y_{ijs(t=3)}$ is a function of participant *i*'s treatment status (where *Coaching*=1 if assigned to receive coaching and *Coaching*=0 if assigned to participate in self-reflection), and individual characteristics (X_{ij}) measured at baseline (including participants' gender, race, high school GPA). The model also controls for study fixed effects (α_s), course section fixed effects (λ_j) and controls for coach and interactor, as well as baseline performance outcome measures. The coefficient β_1 represents the cross-study pooled effect of coaching and is reported in Table 9. Table 9a reports the cross-study pooled effect of coaching with the propensity weights for Study 4.

Assessing robustness of coaching effects. We then extended our meta-analyses by comparing coaching effects across different pairs of studies to understand whether coaching effects are robust across pedagogical task, cohort of teacher candidates, sample and context characteristics and mode of delivery. To assess the replicability and robustness of coaching effects across the systematic sources of variation, we employed distance-based correspondence measures (LaLonde, 1986; Steiner & Wong, 2018). Correspondence measures have traditionally been used to evaluate within-study comparisons (comparing effect estimates from an observational study with those from an RCT benchmark) but can also be used to compare the direction, magnitude and statistical significance patterns of treatment effects across studies. If the difference in treatment effects between two studies is small enough, then we can claim equivalence in the impact of the treatment across the studies. Combining distance-based correspondence measures with conceptual replication study designs allowed us to identify potential sources of treatment effect variation in order to generalize effects for different populations, contexts and settings (Wong et al., 2020). A summary of the correspondence-based measures for each set of comparisons across studies is presented in Table 10.

Results

Baseline Equivalence Across Studies

Tables 5, 7 and 8 reports results of our replication assumptions checks across the five studies. Examining differences in balance across the studies is important to determine all potential sources of coaching effect heterogeneity. If there are significant differences across the studies even before coaching was delivered, these differences could explain potential lack of robustness in coaching effects observed across studies. In

contrast, if there are no significant differences in the studies (other than factors we intentionally vary), we are able to attribute any failure in robustness of coaching effects to specific sources of variation that were systematically altered across studies.

Systematic variation factors. We first examine participant characteristics including GPA, parental education, demographic characteristics (gender, age and race) and high school characteristics (including location, majority race of peers and average achievement level of peers). Table 5 shows that teacher candidate characteristics in studies 1, 2, 3 and 5 are balanced on most covariates, including GPA, parental education, demographic characteristics (including gender, age and race) and high school characteristics. However, undergraduate students in the conceptual replication study (Study 4) differ from the samples of teacher candidates in important ways. First, the teacher candidate samples are significantly more female and older than the undergraduate sample. This is understandable, given that the teacher candidate pool draws from both undergraduate and graduate students who are training to become teachers. Participants in the undergraduate sample also attended high schools with different characteristics than teacher candidates. On average, the high schools that the undergraduate sample attended were more urban, composed of students with higher SES levels and were surrounded by a higher proportion peers who were high achieving. These differences are significant, and they highlight that unit characteristics differ in important ways between the teacher candidate sample and the undergraduate sample.

Appendix Table 3 displays standardized raw differences between the teacher candidate sample in Study 3 and the undergraduate sample in Study 4 prior to and after propensity weighting. Across the covariates, standardized differences in just four (% age

above 21, % of participants who attended an urban high school), overall self-efficacy beliefs and average culturally responsive teaching self-efficacy beliefs fall within the acceptable magnitude of 0.10. All raw standardized differences for the remaining covariates are 0.10 or greater. Using propensity score weighting, all standardized differences of covariates in Appendix Table 3 are less than 0.03 and there acceptable by convention.

We also assess the extent to which the context of the simulation sessions differed across the different studies. Specifically, we are interested in differences in the teaching task that participants completed in the scenario, as well as the training context for these sessions (Table 7). Participants in studies 1, 3, 4 and 5 practiced setting classroom norms and redirecting off-task student avatar behavior in the simulator. Participants in Study 2 practiced providing high quality feedback during a text-based discussion in the simulator. Although the structure of the coaching intervention was consistent across the studies, the focus of the conversation differed based on the teaching task that participants practiced in the simulator.

Another important source of variation is the training context for participants in each of the studies. Participants in studies 1, 2, 3 and 5 were teacher candidates enrolled in the university's teacher preparation program. For these teacher candidates, simulation sessions were integrated into a sequence of methods courses that concurrently discussed practices aligned with the simulator teacher task. In contrast, the undergraduate students in Study 4 were enrolled in a class offered by the school of education structured as an introduction to teaching as a profession. The teaching as a profession course did not cover the management practices targeted in the simulated teaching task, but provided a

more general overview of classroom teaching and historical and cultural aspects of the teaching profession.

Additionally, we are interested in measuring variation across studies in the mode of delivery of simulated sessions. Participants in studies 1, 2, 3 and 4 completed simulated sessions in person in a dedicated simulator lab, whereas participants in Study 5 completed sessions online, over zoom. In-person sessions meant that participants received direct, face-to-face coaching, whereas in practice, online sessions involved coaches providing participants with feedback virtually. Although the coaching protocol remained consistent across the two studies, it is possible that delivery of coaching differed slightly based on the platform used for the simulator sessions.

Systematic replication factors. In addition to assessing balance across the systematic variation factors (which we intentionally varied across studies), we also assess equivalence across factors that we kept constant across the different studies, which we refer to as systematic replication factors. These results are presented in Table 8. First, we examine the extent to which our outcome of interest (overall quality score) remained stable across the studies. Table 8 reports the pre-coaching mean and standard deviation for the overall quality score for each of the four studies. Although the mean pre-test score for undergraduate participants in the conceptual replication is lower than the mean pre-test score for teacher candidates in other studies, these differences are not statistically significant. We also expect to see lower performance scores for the undergraduate sample in Study 4 since they do not have the pedagogical and content training that teacher candidates in the other study samples receive as part of their preparation experience.

Next, we assess whether the coaching intervention was implemented with fidelity across the different studies. To do so, we employ natural language processing techniques to produce a quantitative summary of coaching implementation (see Anglin & Wong, 2020). Table 8 shows that across the five replication studies, adherence to the coaching protocol was relative stable.

Impact of Coaching on Participants' Instructional Practices

Table 9 presents effect size estimates of the impact of coaching on participants' instructional practices for each of the studies. Columns 1-5 provide estimates for each of the experimental replication studies. Table 9a provides estimates for the propensity weighted undergraduate sample in Study 4. Effect size estimates for the overall quality score range from 0.63 SD (for Study 4) to 1.75 SD (for Study 1). From tables 9 and 9a, it appears that there is some heterogeneity in coaching effects across the studies, particularly when we compare studies 1 and 4 both prior to and after propensity weighting (samples of pre-service teacher candidates and undergraduate students respectively).

Meta-analytic Coaching Effect across Studies

A meta-analytic approach pools treatment effects across different studies and presents an overall average treatment effect. We present the meta-analytic coaching effect using a fixed effects model in column 6 of Table 9. In addition to these estimates, forest plots of effect sizes on participants' instructional performance provide visual evidence for our pooled estimates (Figure 7). For each study, a square shows its place on the scale and the confidence interval is represented by the line on either side of the square. Effect sizes for individual studies range from 0.41 SD (for Study 4) to 1.73 SD (for Study 1). Prior to

applying propensity weights to reweight the undergraduate sample in Study 4, the pooled meta-analytic effect across the five replication studies is 1.49 SD[4]. Once we apply propensity weights, the meta-analytic effect across the five replication studies is 1.52 SD. Using Higgin's et al. (2003) recommendation, we calculate the I^2 statistic which represents the percentage of variation attributable to heterogeneity. With our sample of studies, the I^2 statistic of 60% suggests a moderate level of heterogeneity in effect sizes (Higgins, 2003).

Identifying Sources of Coaching Effect Heterogeneity

Given that our meta-analytic results suggest that there is significant heterogeneity in coaching effects across the five replication studies, we compare coaching effects across studies to identify systematic sources of variation and the extent to which our coaching effects are robust across these sources. Traditionally, researchers compare the direction, magnitude and statistical significance patterns of effects across studies to assess replication success (Wong & Steiner, 2018). While these correspondence measures are helpful at understanding *when* treatment effects are/are not robust across studies, they do not provide diagnostic information on *why* treatment effects are/are not robust. This limitation is addressed by using prospective research designs where the research team is able to control the source of variation in a highly standardized simulated setting. In this paper, we use correspondence measures to assess when coaching effects do not replicate across studies, and take advantage of prospective research designs to formally investigate the sources of treatment effect heterogeneity using the causal replication framework and

[4] When we include results from the additional study described in footnote 1, the pooled meta-analytic effect across the six replication studies decreases to 1.18 SD.

its assumptions. A summary of the correspondence measures across each of the replication study comparisons is provided in Table 6.

Pedagogical teaching task. From Table 9 and Figure 7, it appears that providing the same sample of teacher candidates with coaching was similarly effective across the two pedagogical teaching tasks. When practicing setting classroom norms while redirecting off-task behaviors in Study 3, coaching improved candidates' performance by 1.38 SD (column 3 of Table 9). In Study 2, coaching improved teacher candidates' overall quality score by 1.57 SD when the same sample of teacher candidates practiced providing student avatars with high-quality feedback. As summarized in Table 10, both estimates are comparable since they both trend in the same direction (suggesting that coaching led to improvements in candidates' performance) and are both statistically significant at the 99% confidence level. Moreover, the difference between the two coaching effects is not statistically significant ($p>0.10$).

Timing of study. Table 9 and Figure 7 also show that coaching effects are robust across two different cohorts of teacher candidates enrolled in the same teacher preparation program and methods courses. The average coaching effect for the second cohort of teacher candidates (Study 3) suggests that providing coaching improved candidates' overall quality score by 1.38 standard deviations while the corresponding effect for the first cohort of teacher candidates (Study 1) is 1.75 standard deviation units. Using correspondence measures related to the direction, magnitude and statistical significance patterns of coaching effects across the two studies, Table 10 shows that providing pre-service teachers with coaching leads to similar effects when the samples of pre-service teachers are similar on observable characteristics and receiving the same

training. Similar to the pattern observed when comparing coaching effects across two pedagogical teaching tasks, we find that the difference in coaching effects across two cohorts of teacher candidates who practice the same task is not statistically significant ($p>0.10$).

Unit characteristics and settings. Comparing coaching effects from studies 3 and 4 suggest that the impact of providing participants with targeted coaching on their performance in the simulator depends on the sample and setting characteristics of the participants. Table 9 shows that providing a sample of teacher candidates who participate in simulated sessions while simultaneously completing methods courses that focus on classroom management with targeted coaching improved their performance in the simulator by 1.38 standard deviations (Study 3). In comparison, when a sample of undergraduate students completed the same simulated sessions and were provided the same kind of coaching (Study 4), coaching effects were much smaller in magnitude- 0.63 standard deviations (prior to reweighting samples). It is particularly noteworthy that while the coaching effect for the teacher candidates is statistically significant at the 99% confidence level, the coaching effect for the sample of undergraduate students was only marginally significant. As reported in Table 10, the difference in coaching effects between the two studies is significant ($p<0.01$).

Because both samples of participants completed standardized simulated sessions that were designed to be parallel to each other in structure and format, focused on the same pedagogical task (setting classroom norms) and received the same form of coaching, we hypothesize that the observed differences in coaching effects can be attributed to differences in sample characteristics, training context, or a combination of

the two. As evident from Table 5, there are important differences in sample characteristics between the two studies. Once we applied propensity weighting methods to make the sample of undergraduate participants in Study 4 more similar to teacher candidates in Study 3, we observe that there are still significant differences in coaching effects between the two studies. Table 9a shows that after reweighting, teacher candidates who received coaching in Study 3 performed significantly better than their peers who received self-reflection (1.38 standard deviations). For the reweighted undergraduate sample in Study 4, participants who received coaching did perform significantly better than their peers who received self-reflection, but the coaching effects were much smaller in magnitude (0.60 standard deviations). The difference in coaching effects between the two groups remains significant after using propensity score weighting to make the two samples more homogenous (Table 10). The stability of the difference in coaching effects prior to and after reweighting participants in studies 3 and 4 suggests that coaching effects are not across different training contexts, even when participants are similar on observable characteristics

Mode of delivery. Table 9 and Figure 7 also show that coaching effects are comparable in direction, magnitude and significance patterns for two different cohorts of teacher candidates enrolled in the same teacher preparation program and methods courses who participate in two different formats of simulated sessions. The average coaching effect for the first cohort of teacher candidates (Study 3) who completed simulated session in-person (and therefore received face-to-face coaching if they were randomly assigned to the treatment condition) suggests that providing coaching improved candidates' overall quality score by 1.38 standard deviations. The corresponding

coaching effect for the second cohort of teacher candidates (Study 5) who completed simulated sessions online (and therefore received virtual coaching) is 1.64 standard deviation units. The difference in coaching effects across the two studies is not statistically significant ($p>0.10$), as presented in Table 10.

Discussion

The field of teacher preparation needs more evidence on the best ways to efficiently and effectively promote teacher learning through standardized practice opportunities and directive coaching supports. With the shift to practice-based teacher education, preparation programs across the country are increasingly looking for feedback supports that are scalable, targeted and improve candidates' teaching practice (Scheeler, McAfee, Ruhl & Lee, 2006). Used extensively with in-service teachers, coaching has been shown to improve lesson planning and classroom organization, use of instructional skills, teacher attitudes, feelings of self-efficacy and student achievement (Kohler, Ezell & Paluselli, 1999; Guiney, 2001; Neufeld & Roper, 2003). Although nascent, early work in the pre-service context suggests that coaching (in particular, directive coaching) improves teacher candidates' instructional practices (Cohen et al., 2020). Given the resource intensive nature of coaching, however, teacher preparation needs more causal evidence on the efficacy of such coaching supports, as well as on the *context* and *conditions* under which providing teacher candidates with the opportunity to practice their skills and receive targeted, individualized coaching is likely to be effective.

In this paper, we use prospective replication research designs and the Causal Replication Framework to implement five experimental studies that evaluate the impact of targeted, directive coaching on improving teaching skills in a simulated environment.

Across four of the studies, involving multiple cohorts of teacher candidates enrolled in a single, university-based teacher preparation program, we see significant improvements in teacher performance as a result of receiving targeted coaching from expert coaches. This suggests that teachers do not only have to learn in their first few years in the classroom, but they can also benefit from practice opportunities coupled with feedback from experts. Rather than wait until teacher candidates are in student teaching placements, structured practice opportunities and targeted feedback that are integrated into methods coursework can improve teaching practices and prepare teacher candidates for skills with which they often report struggling (Grossman et al., 2009b).

In addition to proving the utility of simulated teaching opportunities in assessing the causal effects of different kinds of interventions on candidates' skills, this study is also one of the first that uses simulated environments to evaluate the robustness of coaching effects across different teaching tasks, study timings, training contexts and modes of delivery of practice and coaching (Cohen et al.,, 2020). Since each study was designed prospectively, the research team systematically introduced variation in each study to examine heterogeneity in observed coaching effects. Our findings suggest that coaching significantly improves participants' teaching skills, and that coaching effects replicate different pedagogical task, multiple cohorts and mode of delivery of simulated session but not across different training contexts, even when participants are similar on observable characteristics.

Our finding that coaching effects are not robust across different training contexts even after reweighting samples of participants so they are more similar on observable characteristics suggests that coaching should be accompanied by instruction that focuses

on evidence-based practices, as well as the opportunity to implement these practices either in a practicum or student teaching setting (Elmore, 2006; Scheeler et al., 2006). That is, practice opportunities should not be stand-alone experiences. This implication is in line with previous studies that highlight the importance of some grounding practice opportunities and coaching with instruction and training (Elmore, 2006; Kraft et al., 2018; Scheeler et al., 2009).

In our studies, teacher candidates who completed simulated sessions were concurrently enrolled in methods coursework focusing on the same content area as the simulator task- i.e., when teacher candidates practiced setting classroom norms in the simulator, they were enrolled in methods courses focusing on classroom management, and when they practiced providing student avatars with high-quality feedback, they were developing the requisite content knowledge skills in their courses. Conversely, participants in the undergraduate sample were enrolled in an introduction to teaching course that was meant to provide students with an overview of teaching as a profession rather than focus on specific content or pedagogical skills. Smaller coaching effects for the undergraduate sample may also be explained by their lack of prior knowledge related to setting classroom norms or managing a classroom. It may be that schema or prior knowledge about the content can support learning from coaching.

An important next step in this work will be to examine whether the pattern of robustness in coaching effects extends to other sites and university contexts, as well as to the more distal outcomes of teaching real children in real classrooms. It is clear from the findings in this paper that coaching can be used to prepare teacher candidates enhance candidates' skills across different classroom scenarios, cohorts, and modes of coaching.

This suggests potential value in integrating coaching more consistently across PBTE coursework. At present, we are in the middle of partnering with other university-based teacher preparation programs to examine the robustness of coaching effects across different populations of candidates, working in diverse geographic locations and classroom settings. Extending the evidence base about the degree to which targeted, directive coaching can improve teaching practices across research sites, samples, and teaching contexts is a critical area for ongoing and future work.

Figure 5

Continuous, Planned Interventions Across Three Academic Years

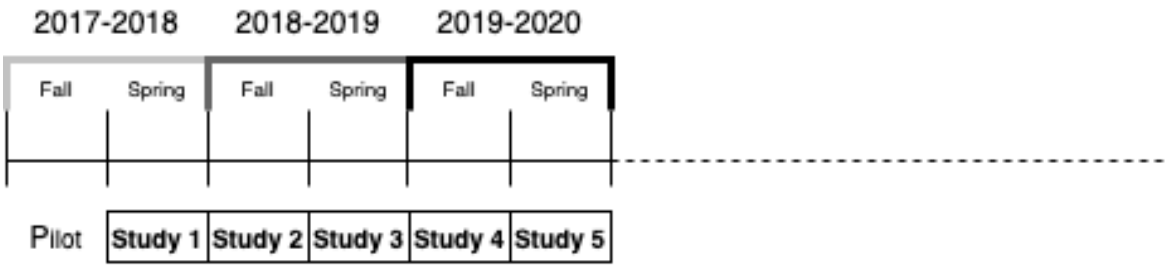
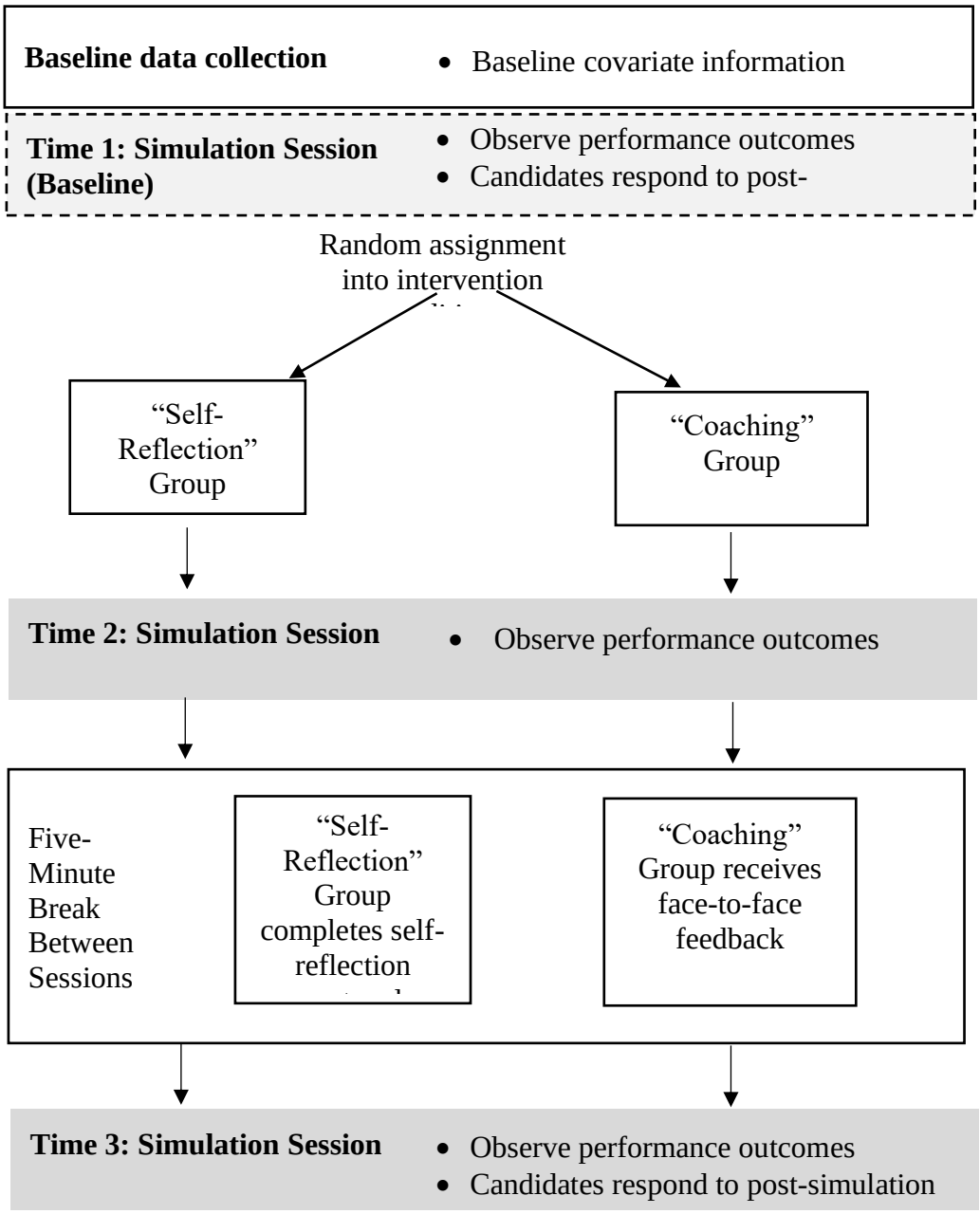


Figure 6

Experimental Design and Study Protocol



Note. Baseline Sessions Completed Only By Teacher Candidates

Figure 7

Forest Plot of Meta-Analytic Coaching Effects (Unweighted Samples)

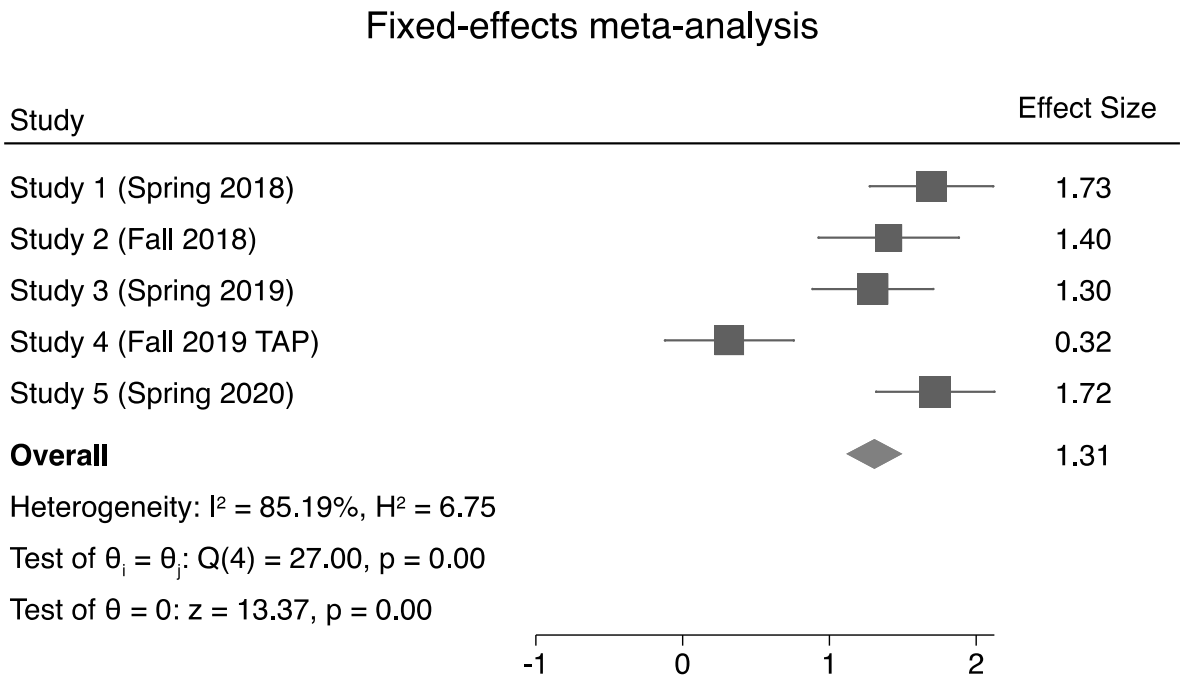


Table 5

Descriptive Statistics Across Replication Studies

	Study 1 (Spring 2018)	Study 2 (Fall 2018)	Study 3 (Spring 2019)	Study 4 (Fall 2019 TAP)	Study 5 (Spring 2020)
	Regression-adjusted Mean				
GPA	3.49	3.54	3.48	3.50	3.57
% Either parent a teacher	0.31	0.44	0.42	0.27	0.44
% Mother education- college or above	0.85	0.98	0.98	0.84	0.84
% Father education- college or above	0.73	0.84	0.85	0.85	0.76
% Female	0.79	0.75	0.73	0.50	0.72
% Over the age of 21	0.47	0.45	0.50	0.36	0.43
% White	0.80	0.85	0.86	0.60	0.73
Location of high school attended					
% Rural	0.18	0.19	0.21	0.13	0.06
% Suburban	0.59	0.51	0.65	0.69	0.20
% Urban	0.22	0.28	0.14	0.18	0.67
Average SES of high school attended					
% Low SES	0.10	0.08	0.08	0.04	0.05
% Middle SES	0.51	0.70	0.70	0.62	0.12
% High SES	0.34	0.22	0.23	0.33	0.54
Majority race of high school attended					
% Primarily students of color	0.08	0.10	0.08	0.08	0.07

% Mixed	0.50	0.55	0.53	0.39	0.23
% Primarily white students	0.35	0.36	0.41	0.52	0.23
Average achievement level of high school attended					
% Primarily low achieving	0.11	0.07	0.07	0.04	0.08
% Primarily middle achieving	0.42	0.57	0.54	0.42	0.42
% Primarily high achieving	0.47	0.36	0.39	0.54	0.50

Notes: Demographic information comes from data collected by the teacher preparation program or administered as surveys to study participants. Each row represents results from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks.

Table 6

Descriptive Statistics for Measures (Including Reliability Alphas)

	Teacher candidate sample			Undergraduate participant sample	
	Range	Mean	Range of reliability alphas	Mean	Reliability alphas
Neo Five-Factor Inventory					
Neuroticism	1-5	2.75	0.85 - 0.89	2.64	0.82
Extraversion	1-5	3.57	0.85 - 0.92	3.35	0.82
Openness	1-5	3.44	0.75 - 0.88	3.07	0.75
Agreeableness	1-5	3.85	0.73 - 0.92	3.06	0.79
Conscientiousness	1-5	3.86	0.85 - 0.95	3.59	0.84
Overall Self-Efficacy	1-9	6.43	0.97 - 0.98	6.24	0.94
Multicultural Attitudes Survey	1-5	4.13	0.88 - 0.90	3.31	0.85
Culturally Responsive Teaching Self-Efficacy	0-100	67.41	0.97 - 0.98	66.85	0.95

Table 7

Systematic Variation Factors Across Studies

	Study 1 (1)	**Study 2** (2)	**Study 3** (3)	**Study 4** (4)	**Study 5** (5)
Level of pedagogical knowledge	Average	Average	Average	Average	Average
Observable characteristics	Similar	Similar	Similar	**Different**	Similar
Pedagogical Task	Setting Norms	**Providing Feedback**	Setting Norms	Setting Norms	Setting Norms
Training Context	Methods Course	Methods Course	Methods Course	**Teaching as a Profession**	Methods Course
Mode of Delivery	In-person	In-person	In-person	In-person	**Virtual**
R.3. Timing	Spring 2018	Fall 2018	**Spring 2019**	Fall 2019	Spring 2020

Table 8

Systematic Replication Factors Across Studies

	Study 1 (1)	**Study 2** (2)	**Study 3** (3)	**Study 4** (4)	**Study 5** (5)
Replication Study Assumptions					
Outcome Stability (Pretest mean & SD)	3.64 (1.22)	3.94 (1.30)	3.46 (1.33)	-	2.82 (0.86)
Coaching (Treatment) Fidelity	0.28	0.44	0.33	0.27	--
Individual Study Assumptions					
Research Design Used	RCT	RCT	RCT	RCT	RCT
Estimation Strategy Used	Robustness checks	Robustness checks	Robustness checks	Robustness checks	Robustness checks
Independent Reproducibility	NA	NA	Yes	NA	NA

Notes: The diagnostics presented here examine the extent to which replication and individual study assumptions under the Causal Replication Framework were met. To assess "outcome stability" across studies, the research team examined the means, standard deviation and measurement properties of the quality performance outcome across studies. "Coaching stability" was assessed using the semantic similarity approach described in Anglin and Wong (2020); a higher score indicates higher similarity to a benchmark scripted treatment protocol. To examine the validity of the RCT, the research team examined baseline equivalence on an array of baseline characteristics for each study. To assess the sensitivity of effect estimates to different model specifications, the research team reports the robustness of results with different control covariates included in the models. All effect estimates were reproduced by an independent analyst with access to the original data and syntax files but was blinded to original study results.

Table 9

Average Treatment Effect by Study and Pooled Meta-Analytic ATE

	Study 1 Mean (1)	**Study 2** Mean (2)	**Study 3** Mean (3)	**Study 4** Mean (4)	**Study 5** Mean (5)	**Meta-analytic** Mean (6)
Overall Quality Score	1.75**	1.57**	1.38**	0.63+	1.64**	1.49**

Notes: Adjusted coaching effects are reported in each column. Coefficients and standard errors in columns (1) through (5) represent standardized mean adjusted differences between control and coaching conditions taken from regressions of the outcome on coaching assignment for each study. Column (6) represents the overall meta-analytic coaching effect. Models include controls for randomization blocks, participants' gender, race, high school GPA, baseline score and interactor and coach fixed-effects.

Table 9a

Average Treatment Effect by Study and Pooled Meta-Analytic ATE for Weighted Study 4 Sample

	Study 1 Mean (1)	**Study 2** Mean (2)	**Study 3** Mean (3)	**Study 4*** Mean (4)	**Study 5** Mean (5)	**Meta-analytic** Mean (6)
Overall Quality Score	1.75**	1.57**	1.38**	0.60+	1.64**	1.52**

Notes: Adjusted coaching effects are reported in each column. Coefficients and standard errors in columns (1) through (5) represent standardized mean adjusted differences between control and coaching conditions taken from regressions of the outcome on coaching assignment for each study. Column (6) represents the overall meta-analytic coaching effect. Models include controls for randomization blocks, participants' gender, race, high school GPA, baseline score and interactor and coach fixed-effects.

Table 10

Summary of Distance-Based Correspondence Measures Across Replication Studies

	N_B	N_R	Source of variation	Distance-based measures			
				Magnitude of effects	Sign of effects	Significance patterns	Difference between coaching effects
Switching replication study (Study 2 and Study 3)	99 ◆	98 ◆	Pedagogical teaching task	✓	✓	✓	-0.19
Multiple cohort study (Study 1 and Study 3)	99	98	Cohort of teacher candidate (time)	✓	✓	✓	-0.37
Conceptual replication study (Study 3 and Study 4)	98	95	Sample characteristics and training context	✓	✓	×	-0.75**
Conceptual replication study (Study 3 and Study 5)	98	104	Mode of delivery	✓	✓	✓	0.15

Notes: ◆ Indicates dependent sample for switching replication design

CHAPTER IV: EXAMINING THE IMPACT OF MENTAL REHEARSAL ON PRE-SERVICE TEACHERS' PERFORMANCE AND PERCEPTIONS IN MIXED-REALITY SIMULATIONS

A large body of literature has documented the difficulties that pre-service teacher candidates face when they first enter the classroom, using terms such as "reality shock", "survival phase" (Huberman, 1989) and "transition shock" (Corcoran, 1981). Novice teachers consistently report feeling underprepared and overwhelmed by teaching responsibilities when they first become teachers of record (Beran, 2005; Mitchell & Arnold, 2004). In particular, managing classroom environments and student behaviors has been often cited as the primary source of stress, anxiety and burnout by new teachers, who are almost twice as likely to leave the profession as their more experienced peers (Brunsting, Sreckovic & Lane, 2014; MacQueen, 2006). Poorly trained teachers are also more likely to use ineffective management methods with students, perpetuating a cycle of student misbehavior and teacher stress that leads to adverse outcomes for both groups (Kaff, Zabel & Milham, 2007).

Even when teacher candidates have the skills necessary to effectively redirect off-task student behavior, they often report feeling emotionally exhausted by the process of managing a classroom environment (Jennings & Greenberg, 2009) on a regular basis. Research suggests that in order to effectively manage disruptive students, novice teachers need a combination of pedagogical knowledge regarding best practices as well as emotional resources to counter the stressful nature of redirecting student behaviors and managing a classroom. That is, when teachers are emotionally exhausted or perceive a

classroom management task as a threat to their efficacy as a teacher rather than as an opportunity to learn, they are less likely to be effective at managing disruptive student behaviors (Seiz et al., 2015). This is in line with cognitive appraisal theory, which highlights than an individual perceives a task or situation as stressful when they determine that the resources or skills they have at their disposal are insufficient to meet the demands of the task (Harvey, Nathens, Bandiera & LeBlanc, 2010; Lazarus & Folkman, 1984). Sutton (2007) finds that teachers often employ cognitive appraisal when faced with disruptive students, and thus recommends that teachers should be equipped with strategies that can help change their perceptions regarding classroom management in addition to more traditional supports that improve knowledge of how to effectively manage students.

One promising set of interventions that may be used to improve performance and perceptions related to classroom management for pre-service and novice teachers are mental rehearsals. Mental rehearsal is a cognitive reappraisal strategy that has been used in healthcare and sports settings to improve performance on a specific task related to the intervention, as well as participants' perceptions of the task including stress, self-efficacy and confidence levels (Arora et al., 2011a; Eldred-Evans et al., 2013). Essentially, mental rehearsal uses a script (or stimulus) that focuses on strengthening the imagery representation of the skills that a participant is expected to use in a particular situation or task (Arora et al., 2011a). By allowing participants to mentally visualize themselves executing a task before performing it, mental rehearsal serves as a preparation technique that improves performance. In addition, the organized format of mental rehearsal scripts reduce cognitive load and benefits learners (particularly novices) by allowing them to

cognitively reappraise the task in the absence of physical exercise, ultimately improving their perceptions of the task.

In this study, I evaluate the efficacy of a mental rehearsal intervention in reducing pre-service teacher candidates' stress levels, improving their management skills and perceptions of student behaviors (including ratings of simulated student behaviors, endorsements of punitive "next-steps" to manage such behaviors and self-efficacy beliefs) during simulated teaching sessions. Implemented as part of a larger experimental study examining the efficacy of providing pre-service teacher candidates with targeted, directive coaching to improve their teaching practice, candidates were randomly assigned to read a mental rehearsal script prior to engaging in a simulated classroom management task. To understand the impact of mental rehearsal both as a stand-alone intervention as well as in combination with more traditional supports offered by preparation programs (such as coaching), I answer three research questions: 1) Does providing pre-service teacher candidates with mental rehearsal improve their perceptions around redirecting off-task student behavior (particularly, their stress levels, self-efficacy beliefs and ratings of student behaviors)? 2) Do candidates who complete mental rehearsal perform significantly better at redirecting off-task student behaviors than candidates who do not? 3) Do the effects of mental rehearsal on candidates' stress levels, self-efficacy beliefs and performance outcomes differ based on whether or not they received targeted coaching supports?

Results suggest that mental rehearsal did not have a significant impact on candidates' self-reported stress levels or self-efficacy beliefs, but significantly improved their performance during the simulated teaching task. Specifically, I find that teacher

candidates assigned to complete mental rehearsal provided more timely redirections (i.e, took lesser time to react) to simulated student avatar behaviors, more specific redirections, more succinct redirections (i.e, took fewer seconds to redirect behaviors), and scored higher on an overall measure of quality of redirections. In addition, I find that while mental rehearsal does not offer candidates who were randomly assigned to receive coaching in-between simulator sessions any advantage (over and above the benefit of receiving feedback from a coach), candidates who were randomly assigned to participate in self-reflection without any guidance from a coach between simulator sessions benefited when they completed mental rehearsal prior to engaging in the simulated task.

This paper proceeds as follows. In Section 1, I provide background and an overview of the literature on the challenges faced by novice teachers around classroom management, the relationship between perceptions and performance in the context of classroom management and the promise of techniques like mental rehearsal in improving performance and perceptions related to specific tasks. In Section 2, I outline research methods, including sample and setting characteristics, experimental design, and my analytic plan. In Section 3, I present results estimating the impact of providing participants with mental rehearsal (and coaching supports) on their stress, self-efficacy beliefs, instructional skills, and perceptions of student behaviors in a simulated teaching context. In Section 4, I discuss my findings, provide implications for teacher preparation policy and directions for future research.

Background

Reality shock, classroom management and novice teachers

Novice teachers experience a “reality shock” when they first enter classrooms and consistently report feeling unprepared for the multitude and often competing demands they are faced with as teachers of record (Atteberry, Loeb & Wyckoff, 2015; Kraft & Papay, 2014). As novice teachers learn to apply their teaching skills in a classroom, most struggle to transition from a learner to a teacher. In part, the lack of readiness among novice teachers can be attributed to the need for structured coursework and integrated opportunities to practice teaching skills during the preparation phase (Grossman & McDonald, 2009b; O’Neill & Stephenson, 2012). Along with this, there is also evidence that that preparation programs do not sufficiently prepare teachers for the high levels of uncertainty, emotion and stress that accompanies teaching (Leung, Chiang & Wong, 2011).

In particular, novice teachers report struggling with managing their classroom, redirecting off-task student behaviors and establishing classroom norms with their students (Beran, 2005; Mitchell & Arnold, 2004). Although most teacher preparation programs include some training on classroom and behavior management, a focus on evidence-based practices and the opportunity to practice management skills do not accompany such training on a consistent basis (Flower, McKenna and Haring, 2016). In addition, even when teacher candidates have the skills necessary to effectively redirect off-task student behavior, they often report feeling emotionally exhausted and stressed by the process of managing their classroom environment (Jennings & Greenberg, 2009). In the context of teaching (and particularly for novice teachers), knowledge of effective classroom management strategies and a self-perceived sense of preparedness related to classroom management is critical to success (Friedman, 2006; Dicke et al., 2014).

Training teachers to feel like they can effectively manage classrooms and concurrently manage their own emotions is important because unprepared teachers report increased stress, lower levels of job satisfaction and higher rates of teacher turnover (Brunsting et al., 2014) resulting from student behavior problems. A large body of literature has highlighted that first-year teachers are more likely to leave the profession (almost 2.5 times more likely) than their experienced counterparts (MacQueen, 2006; Roness, 2011). Most novice teachers who leave the profession cite classroom management and organization as the primary reason for turnover (Chaplain, 2008). Poorly trained teachers who are unable to regulate their own emotions in the classroom are more likely to use ineffective management methods (Feiman-Nemser, 2003; Kaff et al., 2007; Kokkinos, Panayiotu & Davazoglu, 2004), thus perpetuating the cycle of student behavior and stress. Novice teachers in particular are more intolerant of student behaviors (Kaff et al., 2007) and are more likely to recommend special education referrals to students (O'Neill & Stephenson, 2012) than experienced teachers.

Teacher perceptions, performance and cognitive appraisal

A large body of literature has documented the role that formal coursework and training opportunities play in improving pre-service teachers' pedagogical knowledge, as well as perceptions related to classroom management. While coursework and practice opportunities are theorized to improve both performance and perceptions of ability to effectively manage classrooms (as with most teaching domains), empirical work suggests that even after completing traditional components of teacher preparation, pre-service teachers continue to report that they feel only "somewhat prepared" to manage disruptive

behaviors, non-compliance and disorganization in the classroom (Oliver & Reschly, 2007; O'Neill & Stephenson, 2012).

One explanation for why pre-service teachers continue to feel ill-prepared to manage disruptive classroom environments when they enter the classroom may be their perceptions and emotions regarding the stressful nature of classroom management. Prior work has suggested that teachers need high emotion regulation skills in addition to strong pedagogical knowledge to successfully manage challenging student behaviors (Jennings & Greenberg, 2009; Skinner & Beers, 2016). In their paper on the relevance of teachers' emotional resources for successful classroom management, Siez, Voss and Kunter (2015) hypothesize that using pedagogical knowledge to effectively manage a classroom depends on the emotions and perceptions that a teacher has regarding the task of managing a classroom environment. Specifically, they suggest that only when teachers are able to regulate their emotions around classroom management, will they be able to successfully apply knowledge-based strategies. They find that a combination of pedagogical knowledge and low levels of emotional exhaustion are associated with higher ratings of effective classroom management, suggesting that pedagogical knowledge alone may not be sufficient for effective classroom management.

The link between perceptions and performance in teaching contexts has been documented in prior work. The literature suggests that a teacher's self-efficacy beliefs and confidence in their ability to effectively manage a classroom influences their management behaviors, as well as student behavior (Gordon, 2001; Narvaez, Vaydich, Turner & Khmelkov, 2008; Newman-Carlson & Horne, 2004; Rosas & West, 2009). In other domains such as healthcare and the military, studies examining the effects of stress

or anxiety on performance suggest that performance in high-acuity contexts (such as managing a disruptive classroom environment) can be either enhanced or impacted based on how participants perceive the environment (LeBlanc, 2009). Specifically, when participants perceive an upcoming task as a challenge rather than a threat, performance is more successful. As an example, Blaskovich, Seery, Mugridge, Norris and Weisbuch (2004) find that athletes how appraised an upcoming competition as a threat performed significantly worse than their peers who appraised the same as a challenge. This is consistent with findings from a study on public speaking, where people who appraised an upcoming public speaking situation as a challenge outperformed those who perceived the same situation as a threat (Feldman, Cohen, Hamrick & Lepore, 2004).

Whether an individual perceives a situation or task as a challenge or threat is based on their cognitive appraisal of the task. Based on the cognitive appraisal theory (Lazarus & Folkman, 1984), if an individual determines that the resources they have at their disposal are sufficient to meet the demands of the situation, they appraise it as a challenge and recognize the potential for improvements in self-esteem, learning, performance, etc. However, if the individual perceives resources as insufficient to meet the demands of the same situation, then it is appraised as a threat (Chang, 2009; Lazarus, 1966, 2006; Harvey et al., 2010). One promising set of interventions to improve performance and perceptions of specific tasks and situations focus on cognitive reappraisal, or altering individuals' perceptions of demands in comparison to resources. There is some evidence that teachers often employ cognitive reappraisal when faced with disruptive students (Sutton, 2007), suggesting that any intervention that seeks to improve

performance needs to address teacher perceptions regarding demands and resources of the situation as well.

Mental rehearsal to improve performance and perceptions

A popular strategy that has been used in similar high-stakes occupations as teaching, including sports and medicine, is mental rehearsal. In prior studies, mental rehearsal has been shown to improve participants' performance on a specific task related to the mental rehearsal intervention, as well as their perceptions of the task (Arora eta al., 2011b; Eldred-Evans et al., 2013). Mental rehearsal refers to the "cognitive rehearsal of a task in the absence of overt physical movement" (Driskell, Copper & Moran, 1994; pg.481). When engaging in mental rehearsal, participants read a mental rehearsal script (or audio stimulus) that focuses on strengthening the imagery representation of the skills that a participant is expected to use in a particular situation or task (Arora et al., 2011a). Practically, engaging in mental rehearsal allows an individual to mentally visualize themselves executing a set of skills before performing it, as a preparation technique. The mental rehearsal, or visualization, then activates shared neural representations with actually performing the same task (De Lange, Roelofs & Toni, 2008), therefore improving participants' performance. In addition, mental rehearsal scripts are presented as organized, cohesive learning materials that can be used by participants to assimilate information needed to complete the task. This reduces extraneous cognitive load and benefits learners (or novices) in particular, ultimately leading to improved performance and feelings of preparedness related to the task. The goal of mental rehearsal is to eliminate negative perceptions and improve performance by providing participants with a mental map of a specific task prior to engaging in it (Jones & Stuth, 1997).

Early studies on mental rehearsal found that the technique led to significant improvements in performance (Lee, 1990; Perry, 1939; Ross, 1985), leading researchers to conclude the mental rehearsal has the potential to improve performance on specific tasks (Druckman & Swets, 1988; Feltz, Landers & Becker, 1988). In more recent work on the use of mental rehearsal among novice surgeons, researchers find that mental rehearsal is a time-and-cost effective strategy that improves surgeons' knowledge and confidence (Arora et al., 2011b) and reduce stress (Wetzel et., 2011) related to surgical procedures. In a recent study, Arora et al. (2011b) randomly assigned 45 novice surgeons to either participate in mental rehearsal (intervention condition) or an unrelated activity (control condition) prior to a simulated laparoscopic surgery. Participants completed five consecutive days of simulated laparoscopic surgeries, with mental rehearsal (or the unrelated activity) prior to each practice surgery. The authors find that surgeons who read the mental rehearsal script prior to completing the surgeries were significantly less stressed and performed significantly better than their peers in the control group.

In another paper, Ignacio, Scherphier, Dolmans, Rethans & Liaw (2017) use a pre-and-post design to examine whether mental rehearsal was effective in enhancing rescue-and-resuscitate performance and managing stress among a sample of 100 nursing students working with simulated patients. The authors find that although mental rehearsal significantly improved candidates' performance, it did not seem significantly reduce participants' stress levels on an objective measure of stress. However, qualitative data from interviews with participants suggest that engaging in mental rehearsal enabled them to manage their stress and had a calming effect on them prior to engaging in the simulated task.

In all, mental rehearsal interventions show promising results on participants' performance and perceptions in simulated tasks in healthcare settings. Like medical professionals, pre-service and in-service teachers also face challenging and demanding situations every day in the classroom. Along with improving pre-service teacher candidates' skills through coursework, practice opportunities and feedback, teacher preparation programs would be well served by integrating brief supports such as mental rehearsal that help candidates meet the intellectual demands of the classroom, improve their performance and ultimately increase student achievement. The low-cost, low-touch nature of mental rehearsal makes it ideal to implement alongside more traditional supports provided during the preparation phase, including role-plays, rehearsals, simulated teaching opportunities, feedback from mentors, coaching, etc. In fact, Sanders et al. (2008) find that a combination of technical skills training and mental rehearsal is most effective in enhancing performance and improving preparedness related to a specific task.

This study

Although mental rehearsal shows promise in healthcare settings, researchers have not studied the efficacy of mental rehearsal to improve performance and perceptions (such as stress, self-efficacy beliefs, attitudes towards classroom management and endorsements of punitive approaches) in the pre-service teacher preparation context. In this paper, I examine the use of a mental rehearsal script in improving pre-service teachers' outcomes related to classroom management in the context of a simulated teaching opportunity. Specifically, I investigate whether participants who were randomly assigned to read a mental rehearsal script prior to redirecting off-task behaviors and

setting classroom norms with a group of five simulated student avatars performed better, reported lower levels of stress and rated misbehaviors as occurring to a lesser extent than their peers in the control group.

Implemented as part of a larger study that examines the efficacy of providing teacher candidates with coaching supports to improve their practice in the simulator, I combine data from the mental rehearsal intervention with the coaching protocol to understand the impact of mental rehearsal both as a stand-alone intervention and in conjunction with coaching supports (see Cohen et al., 2020 for a more thorough description of the coaching and self-reflection procedures). Candidates randomly assigned to the coaching condition received five minutes of targeted, directive feedback from an expert coach, while candidates randomly assigned to the self-reflection condition completed a self-reflection protocol designed by the research team to reflect on their performance in the simulator. I used a 2x2 factorial design to randomly assign teacher candidates to one of four conditions: *self-reflection with no intervention, self-reflection with mental rehearsal, coaching with no intervention and coaching with mental rehearsal.* Importantly, the mental rehearsal intervention and the coaching intervention were delivered at different times with a five-minute simulation session in-between them, allowing me to estimate the causal impact of mental rehearsal on teacher candidates' outcomes both as a stand-alone intervention as well as when combined with coaching.

Theory of change and research hypotheses

For this study, I theorize that participating in mental rehearsal will improve teacher candidates' performance and perceptions related to classroom management in the context of the simulated practice opportunity. That is, I anticipate that candidates who

complete mental rehearsal will perform better at redirecting off-task student behavior in the simulator, report lower levels of stress prior to their next simulator session, report higher self-efficacy beliefs related to classroom management, and rate student avatar behaviors less harsh than their peers in the control group. The theory of change related to my hypotheses is outlined in Figure 9. Specifically, I use the factorial design implemented in this study evaluates three hypotheses regarding the benefit of mental rehearsal in the context of a mixed-reality classroom management task:

Hypothesis 1. Teacher candidates who participated in mental rehearsal will report lower levels of stress before they complete their first simulator session as compared to control group candidates. That is, candidates who read the mental rehearsal script (regardless of whether they receive coaching or self-reflection after their pre-coaching simulator session) will be less stressed than their peers who do not read the mental rehearsal script prior to their simulator session.

Hypothesis 2. Teacher candidates randomly assigned to participate in mental rehearsal will perform better at redirecting off-task student behavior during their simulator session (prior to participating in self-reflection or receiving targeted coaching) compared to candidates who did not participate in the mental rehearsal intervention. In addition, they will also report more favorable perceptions regarding classroom management (including ratings of student avatar behaviors, endorsements of different classroom management approaches and self-efficacy beliefs related to classroom management). Here, I hypothesize that completing the mental rehearsal exercise (irrespective of whether candidates are assigned to receive coaching or complete the self-reflection protocol following their second simulator session) will improve candidates'

performance in the simulator session and improve their perceptions of the simulated classroom management task by helping them mentally visualize themselves executing a set of skills before performing it, leading to better preparation and better performance.

Hypothesis 3. Teacher candidates who are assigned to coaching/self-reflection (respectively) and mental rehearsal will perform better than participants who were assigned to coaching/self-reflection (respectively) but did not complete the mental rehearsal exercise during their final simulator session when they have completed both mental rehearsal *and* coaching/self-reflection.

Research Methods

Sample and setting characteristics

Designed to complement content and pedagogical knowledge that candidates gained during their methods coursework, the simulated sessions provided candidates with short, standardized, repeated practice opportunities. During a methods course focused on classroom management, teacher candidates were asked to facilitate a classroom discussion about setting classroom norms with five simulated student avatars, while redirecting minor off-task behaviors displayed by the avatars. Behaviors included humming, talking about personal hobbies and minor interruptions, and were meant to inhibit the whole group discussion without being aggressive or overtly defiant. Six off-task behaviors occurred during each five-minute simulated session, providing candidates with the same number of opportunities to redirect behavior. Each teacher candidate participated in three simulator sessions (one at the start of the semester referred to as the first or baseline session, and two back-to-back simulator sessions approximately two months after the first session referred to as the second/third sessions respectively).

Teacher candidates who were randomly assigned to the mental rehearsal condition completed a mental rehearsal script (or protocol) designed by the research team prior to engaging in their second simulator session.

Data included in this study were collected in the Spring of 2020 at a large, selective, public university in the southeast United States that prepares approximately 100 new teachers each year. The overall participant sample was largely White (68%) and female (83%), and middle class (61%). Table 11 provides descriptive statistics of sample characteristics by intervention condition. Baseline covariate measures include candidates' reported characteristics of the high school from which they graduated, parental education, prior academic achievement (grade point average), and scores on self-reported measure of teacher self-efficacy (Tschannen-Moran & Hoy, 2001). Table 11 also includes performance and survey outcome measures at baseline.

Experimental design and data collection procedures

Figure 8 summarizes the experimental design and study protocol for the simulator sessions that candidates completed in Spring 2020. During the baseline period in Figure 8 (Time 1), candidates participated in a baseline simulation session where they were asked to lead a discussion on setting classroom norms while simultaneously redirecting off-task behaviors from student avatars. At this time, the research team collected data on participants' covariate information as well as pretest measures of their performance and perception outcomes. Candidates were then randomly assigned within their elementary or secondary course sections to one of four conditions: Control group with self-reflection (n=27), Mental rehearsal group with self-reflection (n=29), Control group with coaching (n=24) and Mental rehearsal group with coaching (n=26). Although coaches and

simulation specialists were not randomly assigned to sessions, they were scheduled in a manner that allowed sufficient variation across course sections, day and timings of sessions.

After a few months, candidates participated in their first follow-up simulator session at Time 2. Before they participated in the session, all participants were sent a pre-simulator survey that asked them to list the kinds of misbehaviors they anticipated the student avatars would display, how they planned to respond to them, how stressed they felt about their upcoming simulator session and a measure of self-efficacy related to classroom management classroom management. Candidates who were assigned to complete the mental rehearsal intervention received a version of the same survey that contained all of the questions included in the control group survey as well as a mental rehearsal script (script is provided in Appendix B). The follow-up 1 session was designed to be "parallel" to the session at baseline in that candidates were asked to complete the same task of setting classroom norms and redirecting off-task student avatar behavior, but were exposed to different off-task behaviors in each session. The number, type and severity of off-task behaviors were similar and standardized across all sessions. This allowed candidates to approximate redirecting behaviors in similar but not identical settings. Interactors who controlled the student avatars were trained rigorously by the research team to ensure that simulation sessions were delivered with fidelity across teacher candidates. Teaching outcomes were measured for all participants during the follow-up simulator session.

Immediately after their first follow-up simulator session, candidates who were assigned to receive coaching (either with or without mental rehearsal) received five

minutes of targeted coaching from an expert coach. At the same time, candidates who were assigned to self-reflection (either with or without mental rehearsal) responses to a series of reflection prompts designed by the research team without any support from a coach. After the five minutes of coaching or self-reflection, all candidates participated in the third and final simulation session (follow-up 2), during which teaching outcomes were measured once again. After completing the second follow-up simulator session at Time 3, candidates completed post-simulation survey asking them to assess avatar behaviors and recommend "next-steps" for addressing student avatar behaviors if they occurred in a classroom setting, as well as another administration of the self-efficacy in classroom management measure.

Treatment and control conditions

Self-reflection with no intervention. Teacher candidates who were randomly assigned to the self-reflection and no intervention condition did not complete the mental rehearsal exercise. In between two five-minute simulator sessions, these candidates were asked to reflect for five minutes using a self-reflection protocol designed based on the extant literature on teacher self-reflection (Yost, 2006). The protocol included three questions: "What are some things you think went well in terms of redirecting student behavior?" "What are some things you think could have gone differently in terms of redirecting student behavior?" and "What are you going to work on in the next five-min session to improve your redirections of student behavior?"

Self-reflection with mental rehearsal. Prior to their second simulator session, candidates assigned to the self-reflection with mental rehearsal read a mental rehearsal script as part of their pre-simulator survey. They then completed one five-minute

simulator session, the self-reflection protocol described above, and then another five-minute simulator session.

Coaching with no intervention. Teacher candidates randomly assigned to receive coaching with no intervention did not read the mental rehearsal script as part of their pre-simulator survey. Once they completed a five-minute simulator session, candidates received targeted feedback from an expert coach on their performance. Coaches were trained doctoral students who were enrolled at the same university. Coaching conversations followed a similar structure in that coaches would first gauge candidates' understanding of their performance, identify elements of effective instruction, highlight a specific skill in need for improvement, and finally reinforce the link between specific practices and the larger instructional goal. The teacher candidate then had another opportunity to practice their skills in the simulator immediately following their conversation with the coach.

Coaching with mental rehearsal. In the coaching with mental rehearsal condition, teacher candidates read the mental rehearsal script as part of their pre-simulator survey prior to completing the second simulator session. Following this session, they received targeted feedback from an expert coach in the manner described above. All teacher candidates then completed a third and final five-minute session in the simulator.

Outcome measures

In this study, I use data from observational rubrics of teacher candidates' performance in the simulator as well as their responses on pre-simulator and post-simulator surveys (to measure perceptions regarding classroom management). Results

from factor analyses and reliability calculations for the construct measures used in this study are presented in Table 12.

Stress measure. Participants' subjective stress levels were measured using the short form of the State-Trait Anxiety Inventory (STAI). Originally developed by Spielberger, Gorsuch and Lushene in 1970, Marteau and Bekker (1992) created a six-item short form of the STAI that is used in this study. The STAI is a self-report questionnaire that measures participants' subjective levels of stress about an imminent task or situation. It contains 6 items, each assessed on a 4-point scale and totaled to produce an overall score (with a reliability of 0.85). Responses closer to the minimum overall score of 6 equates to the lowest stress, while responses closer to the maximum of 24 suggest higher stress. The STAI has been widely used to measure stress in the psychological literature, including among samples of pre-service teachers (see Topoglu, 2014; Yokus, 2012). In this study, pre-service teacher candidates were asked to complete the STAI as part of the pre-survey that was administered prior to the first follow-up simulator sessions (and immediately following the mental rehearsal intervention) at Time 2 in Figure 8.

Self-efficacy beliefs. Teacher candidates completed the Teacher Sense of Efficacy Scale (TSES) developed by Tschannen-Moran and Hoy in 2001 during the baseline survey at Time 1, pre-survey administered at Time 2, and after the follow-up 2 simulator session at Time 3. The long-form version of the instrument consists of 24 items that are classified into three composite subscales: *1) efficacy for instructional strategies, 2) efficacy for student engagement and 3) efficacy for classroom management*. In this study, I analyze data from the classroom management subscale, which consists of eight

items including “How much can you do to control disruptive behavior in the classroom?”, “How well can you establish a classroom management system with each group of students?” “To what extent can you make your expectation clear about student behavior?”, etc. Responses range from 1 (None at all)-9 (A great deal) on a Likert scale with anchors at 1=Nothing, 3=Very little, 5=Some influence, 7=Quite a bit and 9=A great deal. Reliability for the TSES measure was between 0.93 and 0.94 across the three time points that data were collected in this study.

Candidate Performance in the Simulator. An observational rubric measuring candidates’ ability to redirect off-task behavior according to the Responsive Classroom (2014) framework was developed by the research team. A team of trained coders coded videos for each participant from each simulator session (baseline, follow-up 1 and follow-up 2) for a total of 322 videos. Using Krippendorff’s alpha (1970) to calculate inter-rater reliability, a total of 92 videos were double-coded resulting in overall reliability of 0.73. Coder drift was addressed using weekly calibration checks and agreement reports run by the research team, and scores from double-coded videos were averaged to form aggregate measures of performance. Each video was coded for a) the timeliness of redirections (how quickly the candidate responded to an off-task behavior), b) the proportion of specific redirections provided by the teacher candidate, c) the succinctness of redirections (measured as the number of seconds taken by the teacher candidate to successfully redirect the off-task student behavior) as well as an overall measure of quality of redirections provided by the candidate (on a scale of 1-10).

Ratings of student behaviors and management approach endorsements. In addition to coding videos for candidates’ performance, the research team administered a

post-simulation survey at Time 1 and Time 3 that asked candidates to rate student avatar behaviors using a modified IOWA Connor's rating scale (Waschbush & Willoughby, 2007), and the extent to which they would employ punitive approaches to address these behaviors. The avatar behaviors range from minor off-task behaviors to more extreme behavior. Responses range from (1) "Not at all" to (4) "Very much" for each behavior and were averaged to form a scale with reliability of 0.90. The items about punitive "next steps" asked candidates to rate the extent to which they would endorse different punitive management approaches (including special education and disciplinary referrals) to address behaviors that student avatars exhibited during the simulation session. Responses ranged from (1) "Extremely unlikely" to (10) "Extremely likely" and were averaged to create an endorsement rating as presented in tables 11 and 12.

Analytic Model

Impact of mental rehearsal on candidates' performance and perception outcomes. To examine the average treatment effect (ATE) of providing candidates with a brief mental rehearsal on their stress levels, self-efficacy beliefs and performance in the simulator during follow-up 1 (Time 2), I begin by estimating the following cross-sectional model[5]:

$$Y_{ij(t=2)} = \beta_0 + \beta_1 MentalRehearsal_{ij} + X_{ij} + \lambda_j + \epsilon_{ij} \quad (9)$$

Here, $Y_{ij(t=2)}$ represents the outcome for teacher candidate *i* in course section *j* at during the first follow-up simulator session (prior to participating in self-reflection or

[5] In addition to regression models, I also estimated the impact of mental rehearsal on teacher candidates' performance and perceptions outcomes at Time 2 and Time 3 using Structural Equation Modeling. SEM models present an advantage over regression models in their ability to model measurement errors and variances, while simultaneously allowing assessment of model fit. The analytic model and results from the SEM are presented in Appendix C, and suggest that results are robust to measurement error.

receiving targeted coaching). $MentalRehearsal_{ij}$ is an indicator variable coded to 1 if teacher candidate i was randomly assigned to read the mental rehearsal script and 0 otherwise. In my preferred model specification, X_{ij} includes controls for baseline characteristics of teacher candidates (including their mother's education level, gender, race and indicators for missing information), their baseline performance outcomes in the simulator at Time 1, the interactor who controlled the simulation session at Time 2, and whether the candidate was randomly assigned to the coaching or self-reflection condition. Since candidates were randomly assigned to one of four conditions within course sections, my model includes a series of fixed effects for each randomization block (λ_j). In Equation 9, β_1 represents the adjusted average treatment effect of completing the mental rehearsal protocol relative to the control group during follow-up 1.

Since there may be imbalances between the mental rehearsal and control group by chance, Appendix Table 4 shows the sensitivity of the effect estimates to different model specifications. I estimate treatment effects using five different model specifications- in Model 1 presented in Column 2 of Appendix Table 4, I control only for randomization blocks (λ_j). Model 2 presented in Column 3 adds an indicator for whether teacher candidates were randomly assigned to the self-reflection or coaching condition, Model 3 in Column 4 controls for baseline characteristics of teacher candidates (including their mother's education level, gender, race and indicators for missing information). Model 4 in Column 5 adds candidates' baseline performance outcomes in the simulator at baseline while Model 5 in Column 6 of Appendix Table 4 controls for variation in the interactor and coach who controlled/managed simulator sessions during the follow-up period. Overall, I find that the magnitude of effects is robust across model specification.

In addition to the cross-sectional models presented in Equation 9, I also examine results from longitudinal models to understand how mental rehearsal impacts candidates' outcomes in the simulator during follow-up 2. In the model presented below, I examine changes in teacher candidates' outcomes (both simulator performance and self-efficacy beliefs in classroom management) by mental rehearsal condition over time. Here, I hypothesize that participants who completed the mental rehearsal protocol prior to simulation sessions at the first follow-up perform better than their peers who were assigned to the control condition and did not complete the mental rehearsal (after controlling for whether or not participants were randomly assigned to self-reflection or coaching). To examine this hypothesis, I estimate the following model[1]:

$$
\begin{aligned}
& Y_{ijt} \\
& = \beta_0 + \beta_1 MentalRehearsal_{ijt} + \beta_2 Time2_t + \beta_3 Time3_t + \beta_4 Coaching_{ijt} \\
& + \beta_5 MentalRehearsal_{ijt} * Time2_t + \beta_6 MentalRehearsal_{ijt} * Time3_t \\
& + \beta_7 MentalRehearsal_{ijt} * Coaching_{ijt} + \beta_8 Coaching_{ijt} * Time2_t \\
& + \beta_9 Coaching_{ijt} * Time3_t + \beta_{10} MentalRehearsal_{ijt} * Coaching_{ijt} * Time2_t \\
& + \beta_{11} MentalRehearsal_{ijt} * Coaching_{ijt} * Time3_t + X_{ij} + \lambda_j \\
& + \epsilon_{ij} \qquad (10)
\end{aligned}
$$

In the equation above, Y_{ijt} represents average outcomes for teacher candidate *i* enrolled in course section *j* at Time *t* (coded as 1 for baseline, 2 for follow-up 1 and 3 for follow-up 2). β_1 represents the average difference between the mental rehearsal group and the control group in Y during the baseline simulator session. $\beta_1 + \beta_2$ represents the average difference between outcome *Y* for mental rehearsal and control groups at follow-up 1 (Time 2), while $\beta_1 + \beta_6$ represents the difference between outcome *Y* for the mental

rehearsal and control groups at follow-up 2 (Time 3). X_{ij} includes controls for teacher candidate characteristics as well as baseline outcome measures and controls for whether the teacher candidate was randomly received coaching between simulator sessions at Time 2 and Time 3. λ_j represents fixed effects for course section and ϵ_{ij} is an error term. Results from these models are presented in Table 13.

Impact of mental rehearsal on candidates' performance and self-efficacy beliefs based on coaching condition. To better understand how coaching supports interacted with mental rehearsal condition to improve teacher candidates' performance and perceptions in the simulated setting, I examine changes in candidates' outcomes by mental rehearsal *and* coaching condition over time. Specifically, I estimate the following longitudinal model:

$$
\begin{aligned}
& Y_{ijt} \\
& = \beta_0 + \beta_1 MentalRehearsal_{ijt} + \beta_2 Time2_t + \beta_3 Time3_t + \beta_4 Coaching_{ijt} \\
& + \beta_5 MentalRehearsal_{ijt} * Time2_t + \beta_6 MentalRehearsal_{ijt} * Time3_t \\
& + \beta_7 MentalRehearsal_{ijt} * Coaching_{ijt} + \beta_8 Coaching_{ijt} * Time2_t \\
& + \beta_9 Coaching_{ijt} * Time3_t + \beta_{10} MentalRehearsal_{ijt} * Coaching_{ijt} \\
& * Time2_t + \beta_{11} MentalRehearsal_{ijt} * Coaching_{ijt} * Time3_t + X_{ij} + \lambda_j \\
& + \epsilon_{ij} \qquad (11)
\end{aligned}
$$

In the equation above, Y_{ijt} represents average outcomes for teacher candidate *i* enrolled in course section *j* at Time *t* (coded as 1 for pre-intervention, 2 for post-intervention/pre-coaching and 3 for post-intervention, post-coaching). β_1 represents the average difference between the mental rehearsal, self-reflection group and the control, self-reflection group in Y at baseline, while $\beta_1 + \beta_7$ represents the average difference

between the mental rehearsal, coaching group and the control, coaching group at the same time point. $\beta_1 + \beta_5$ represents the difference between the average difference between the mental rehearsal, self-reflection group and the control, self-reflection group in Y during follow-up 1 at Time 2, and $\beta_1 + \beta_5 + \beta_{10}$ represents the difference between the average difference between the mental rehearsal, coaching group and the control, coaching group in Y at the same time point. Finally, $\beta_1 + \beta_6$ represents the difference between the average difference between the mental rehearsal, self-reflection group and the control, self-reflection group in Y during follow-up 2 at Time 3, and $\beta_1 + \beta_6 + \beta_{11}$ represents the difference between the average difference between the mental rehearsal, coaching group and the control, coaching group in Y at the same time point. X_{ij} includes controls for teacher candidate characteristics. λ_j represents fixed effects for course section and ϵ_{ij} is an error term. Results from these models are presented in Table 14.

Results

Stress. Column 2 of Table 13 summarizes the impact of mental rehearsal on teacher candidates' stress level during follow-up 1 at Time 2 (prior to participating in self-reflection or coaching). I find that teacher candidates who were randomly assigned to complete the mental rehearsal protocol report higher levels of stress measured by the STAI as compared to their peers who did not complete the mental rehearsal protocol. Specifically, Table 13 shows that on average, candidates who completed the mental rehearsal protocol scored 0.53 points (0.06 SD) more on the STAI measure as compared to their peers. This difference is not statistically significant ($p>0.10$).

Self-efficacy beliefs. Column 2 of Table 13 also shows teacher candidates' ratings of their efficacy beliefs in managing a classroom immediately after completing

the mental rehearsal protocol at follow-up 1. Coefficients in Column 2 suggest that teacher candidates who were randomly assigned to complete the mental rehearsal protocol report lower self-efficacy beliefs than their peers who did not complete the mental rehearsal protocol. On average, candidates who completed the mental rehearsal protocol scored 0.30 points (0.44 SD) less on the TSES measure as compared to their peers. Although this difference is large in magnitude, differences between the two group are not statistically significant ($p>0.10$).

Performance outcomes. Next, I estimate the impact of mental rehearsal on candidates' performance in the simulator during follow-up 1 (immediately after completing the mental rehearsal protocol, prior to receiving coaching or self-reflection). Specifically, I am interested in whether mental rehearsal improves the specificity, succinctness and timeliness of their redirections, ultimately leading to higher overall quality scores. From Column 2 in Table 13, I find that candidates who were randomly assigned to complete the mental rehearsal protocol performed significantly better than their peers in the control group.

First, candidates in the mental rehearsal group provided more timely redirections in response to off-task behaviors displayed by student avatars. On average, candidates in this group reacted approximately 4 seconds faster than candidates in the control group (β=-4.20, $p<0.01$). Candidates in the intervention group also provided more succinct redirections in response to off-task behaviors- that is, they took lesser time to effectively redirect off-task student avatar behavior. Candidates who completed mental rehearsal took 3 seconds lesser (β=-2.62, $p<0.10$), on average, to redirect off-task behaviors compared to their peers in the control group. Finally, mental rehearsal also improved the

proportion of specific redirections that teacher candidates provided. Candidates in the mental rehearsal group provided more specific redirections than their peers in the control group (β=0.18, p<0.01).

Column 2 in Table 13 also shows a significant difference between the mental rehearsal and control groups on a measure of the overall quality score. This is not surprising, since the overall quality score is a measure of the extent to which teacher candidates provided timely, specific and succinct redirections. Results from Table 13 suggest that candidates who completed the mental rehearsal intervention scored 0.81 points (p<0.01) higher on average than their peers in the control group.

Behavior ratings and endorsements of management approaches. Table 13 reports the impact of mental rehearsal on teacher candidates' ratings of student avatar behaviors and endorsements of punitive management approaches after the simulator session at follow-up 2 (after they have received self-reflection or coaching). Table 13 shows that on average, candidates who completed the mental rehearsal protocol rated oppositional or defiant behaviors as occurring to a significantly lesser extent (β=-0.48, p<0.05) than their peers who did not complete the mental rehearsal protocol. I also find that candidates who completed mental rehearsal were less likely to endorse punitive management approaches (such as disciplinary or special education referrals) to address the off-task behavior as compared to their peers, although this difference is not statistically significant (β=-0.30, p>0.10). An important note is that since perceptions outcomes were measured at follow-up 2, after teacher candidates had either participated in self-reflection or coaching, the estimates of the impact of mental rehearsal on these outcomes are confounded by coaching condition assignment (however, since teacher

candidates were randomly assigned to self-reflection or coaching, I do not anticipate significant changes in the pattern of results observed).

Exploring Dosage Effects Over Time

To understand whether mental rehearsal continued to have an impact on participants' performance and perceptions in the simulator beyond follow-up 1, I examine results from the longitudinal model in Equation 11. Results are presented in Columns 2 and 3 of Table 13, as well as Figures 10 (for overall quality score) and Figure 11 (for self-efficacy beliefs).

Performance outcomes. Figure 10 and Table 13 shows that participating in mental rehearsal prior to the simulator session at Time 2 improves candidates' overall performance at redirecting off-task student avatar behavior. That is, from baseline to follow-up 1, the overall quality score increased from 2.70 points to 4.58 points for the mental rehearsal group compared to an increase from 2.95 points to 3.78 points for the control group. Although both groups saw significant improvements in their overall quality score between the two time points, the increase is greater in magnitude for the participants who completed the mental rehearsal protocol (β=-0.81, p<0.01). This makes sense, given that participants who were randomly assigned to the mental rehearsal condition completed the protocol just prior to the follow-up 1 simulation session at Time 2. From follow-up 1 to follow-up 2 however, both groups saw steep gains in the overall quality score (β=1.18, p<0.01 for the mental rehearsal group; β=1.67, p<0.01 for the control group), likely driven by the fact that participants in both the control group and the mental rehearsal group received coaching from an expert coach between simulation sessions at Time 2 and Time 3. The difference in the gains across both groups is not

statistically significant (β=0.32, p>0.10), suggesting that mental rehearsal provided little prolonged benefit after initial gains observed at follow-up 1.

Table 13 and Figures C4-C6 show similar patterns of results for different outcomes (timeliness of redirections, proportion of specific redirections and succinctness of redirections). Table 13 shows that participants in the mental rehearsal condition provided significantly more timely redirections in response to student avatar behaviors at Time 2 after completing the mental rehearsal in comparison to their peers in the control group. Specifically, participants in the mental rehearsal condition reacted approximately 4 seconds faster to student avatar behaviors at follow-up 1 (with an average of 10 seconds) than they did at baseline (an average 15 points) and the difference between the two time-points is statistically significant (p<0.01). In comparison, participants in the control group did not see much change in their reaction times (β=-0.20, p>0.10). The difference in reaction times between the two groups is statistically significant at follow-up 1 (β=-4.20, p<0.01) as well as follow-up 2 (β=-2.201, p<0.01). Unlike the overall quality score, differences between the two groups is significant for timeliness in redirections (p<0.01) suggesting that mental rehearsal does reduce average reaction times over and above improvements due to coaching.

Similarly, participants in the mental rehearsal condition significantly outperformed their peers in the control group at follow-up 1 for both the proportion of specific redirections provided (β=0.18, p<0.01), as well as the succinctness of redirections (β=-2.62, p<0.10). At follow-up 2, however, differences between the two groups were not statistically significant, although participants in the mental rehearsal group continued to perform better than those in the control group.

Table 13 and Figure 11 however, show that while candidates in both the mental rehearsal group and the control group did experience slight increases in their self-efficacy for classroom management between baseline and follow-up 1, these increases were not significant for either group. Additionally, average self-efficacy beliefs for participants in the mental rehearsal condition were not significantly different from average self-efficacy beliefs for participants in the control group (β=-0.30, p>0.10). At follow-up 2, participants in the mental rehearsal condition reported significantly lower self-efficacy beliefs than their peers in the control group (β=-0.39, p<0.10), suggesting that mental rehearsal may have actually reduced participants' self-efficacy beliefs even more between follow-up 1 and follow-up 2.

Exploring the Impact of Coaching and Mental Rehearsal Over Time

To examine my hypothesis that participants who received mental rehearsal and coaching would perform participants who received only coaching (and participants who received mental rehearsal and self-reflection would outperform those who received only self-reflection), I examine longitudinal models presented in Equation 11, where mental rehearsal condition and coaching/self-reflection conditions were interacted. Results from these models are presented in Table 14 and Figures 12, 13 and C7-C9.

Overall, results presented in Table 14 suggest that participants' responses to mental rehearsal differ based on whether they were randomly assigned to participate in self-reflection or receive targeted coaching between simulator sessions at follow-up 1 and follow-up 2. At follow-up 2, Table 14 and Figure 12 shows that among participants who were randomly assigned to complete self-reflection without any additional support, teacher candidates who completed mental rehearsal scored higher on a measure of overall

quality of redirections (5.13 points) than candidates who did not (4.42 points), and differences between the two groups are statistically significant (β=0.70, p<0.05). In comparison, among participants who were randomly assigned to receive five minutes of targeted feedback from a coach in-between simulator sessions at follow-up 1 and follow-up 2, those who completed mental rehearsal scored slightly lower (6.89 points) than those who did not (6.93) points, although these differences are not statistically significant (β=-0.04, p>0.10).

Figures C7-C9 in the Appendices and Table 14 show a similar pattern for the other coded simulator performance outcomes (timeliness of redirections, proportion of specific redirections and succinctness of redirections). Among those randomly assigned to complete the self-reflection protocol between simulator sessions at follow-up 1 and follow-up 2, those who completed the mental rehearsal protocol outperformed those in the control group at follow-up 2. Specifically, participants in the mental rehearsal, self-reflection group provided significantly more timely redirections (β=-3.63, p<0.01), provided a greater proportion of specific redirections (β=0.08, p>0.10), and provided significantly more succinct redirections (β=-5.69, p<0.01), compared to their peers in the control group. In comparison, among participants randomly assigned to receive five-minutes of targeted feedback from an expert coach in-between simulator sessions at follow-up 1 and follow-up 2, outcomes for those who completed the mental rehearsal protocol did not differ much from participants who did not complete the mental rehearsal protocol. Control, coaching group participants provided less timely redirections (β=-0.07, p>0.10), a slightly greater proportion of specific redirections (β=0.01, p>0.10), and less succinct redirections (β=-0.01, p>0.10) than their peers in the mental rehearsal, coaching

group, although none of the differences between the two groups are statistically significant

Finally, it appears that teacher candidates' self-efficacy beliefs related to classroom management were affected differently based on the combination of mental rehearsal and coaching condition that they were randomly assigned to. Specifically, I find that for participants who were randomly assigned to self-reflection, average self-efficacy beliefs for participants who completed mental rehearsal at follow-up 2 were significantly lower (β=-0.64, p<0.05) than self-efficacy beliefs for participants in the control group at the same time point. Among participants who were randomly assigned to coaching, those who completed mental rehearsal reported slightly lower self-efficacy beliefs than those who did not complete mental rehearsal (β=-0.11, p>0.10) although differences between the two groups are not statistically significant.

Discussion

With novice teachers continuing to report feeling underprepared, overwhelmed and stressed by teaching when they first enter the classroom environment, teacher preparation programs need more empirical literature on the kinds of supports that can be used to improve performance and perceptions regarding teaching during the preparation phase. This is particularly true in the context of classroom management, since it is an area that pre-service and novice teachers report feeling the most ill-prepared for (Beran, 2005; Mitchel & Arnold, 2004). The literature points to the importance of both pedagogical knowledge and emotional resources for effective classroom management, suggesting that any supports during the preparation phase should aim to improve both simultaneously.

This is critical, given that a large body of work documents the relationship between perceptions and performance in high-acuity contexts such as classroom management.

In this paper, I examine the use of a specific cognitive reappraisal strategy called mental rehearsal in improving pre-service teacher candidates' outcomes in the context of simulated teaching tasks around setting classroom norms and redirecting off-task behavior. Specifically, I examine whether teacher candidates who were randomly assigned to participate in mental rehearsal report improved perceptions (lower levels of stress, higher self-efficacy beliefs, and more favorable perceptions of student behaviors) and are more effective at setting classroom norms and redirecting off-task behaviors as compared to their peers in the control group. Because this study was implemented as part of a larger study examining the efficacy of coaching supports on teacher candidates' instructional skills in the simulator, I examine the impact of mental rehearsal both as a stand-alone intervention as well as combined with more traditional supports offered during the preparation phase (such as coaching).

Findings suggest that mental rehearsal did not significantly improve participants' stress levels or self-efficacy beliefs related to classroom management, but did improve their performance in the simulator. Specifically, I find that participants who were randomly assigned to the mental rehearsal condition reacted significantly faster to off-task student avatar behaviors, provided a greater proportion of specific redirections, took fewer seconds on average to effectively redirect behaviors, and scored significantly higher on an overall measure of quality of redirections. These findings corroborate findings from other studies that mental rehearsal results in significant performance improvements during complex simulated tasks (Arora et al., 2011b; Wetzel et al., 2011).

In addition to improved performance, I also find that participants who completed mental rehearsal rated misbehaviors during the simulator session as occurring to a significantly lower extent than participants who did not complete mental rehearsal. Mental rehearsal participants also endorsed punitive management approaches to address off-task behaviors to a lesser extent than their peers in the control group. Taken together, the findings from this study suggest that mental rehearsal significantly improved pre-service teacher candidates' performance at redirecting off task behaviors during the simulator session, as well as their general perceptions regarding classroom management, but did not reduce stress or self-efficacy beliefs among teacher candidates.

Limitations and Future Research

One possible explanation for the lack of impact on teacher candidates' perceptions regarding the classroom management simulation task is that most studies on mental rehearsal examine the impact of repeated mental rehearsals and practice opportunities on participants' stress levels. In the Arora et al. study (2011a), participants completed five rounds of mental rehearsal and simulated laparoscopic surgeries. Their performance at the simulated task was averaged across the five sessions. Thus, the significant reductions in stress that the authors find are likely because participants completed multiple sessions of mental rehearsal and practice sessions. Thus, it is possible that for mental rehearsal to significantly reduce stress levels, participants need to engage in mental rehearsal a few times. In addition, previous work has suggested that self-report measures may not accurately capture reductions in stress levels, particularly before participants have had an opportunity to complete the task that is targeted through mental rehearsal. Further research examining the efficacy of mental rehearsal in improving pre-

service teachers' perceptions around a high-stress task such as classroom management should consider using physiological measures of stress including cortisol samples, heart-rate monitors, etc.

In addition to estimating the impact of mental rehearsal on participants' outcomes as a stand-alone intervention, I examine whether the impact of mental rehearsal differs based on whether or not teacher candidates were randomly assigned to receive coaching supports between their simulator sessions. Results suggest that mental rehearsal offered no added benefit over and above coaching, but for participants who were randomly assigned to complete a self-reflection protocol without additional support, mental rehearsal improved outcomes. Although pilot in nature, findings from this study suggest that mental rehearsal can be a promising tool to improve performance and perceptions related to classroom management during the preparation phase (even in the absence of more traditional supports such as coaching that are used by preparation programs to improve pedagogical skills). Future research should examine the contrast between mental rehearsal and coaching to understand whether a low-cost, low-touch intervention can be used in place of more resource-intensive supports such as coaching.

Conclusion

Overall, this study provides important evidence on the efficacy of cognitive appraisal strategies such as mental rehearsal in improving performance and perceptions related to classroom management, which is an area that novice teachers consistently report feeling underprepared and overwhelmed by. Teacher preparation needs more empirical evidence on what supports can be used to move some of the "on-the-job" learning to the preparation phase. By examining the efficacy of mental rehearsal both as

an individual support as well as in conjunction with coaching, this paper provides useful evidence that preparation programs can use to integrate cognitive appraisal strategies into more complex teaching tasks that pre-service and novice teachers often report struggling with. Additionally, this study leverages simulated teaching practice opportunities, which is a growing area of research in the pre-service teacher preparation context. More research is needed to understand the mechanisms by which mental rehearsal improves performance, perceptions and a combination of the two. Although it is beyond the scope of this experimental study, future work should qualitatively examine whether mental rehearsal improves feelings of preparedness to redirect off-task student behavior, or whether teacher candidates felt less anxious about their upcoming simulator session after completing mental rehearsal. It is clear from these results that mental rehearsal does not improve stress or self-efficacy beliefs in the short team, but it is possible that repeated mental rehearsal practice can significantly reduce stress and improve feelings of preparedness and readiness to tackle a disruptive classroom environment. This study provides promising evidence on the role of mental rehearsal in teacher preparation, and provides directions for future work on examining how cognitive strategies can be used to improve performance and perceptions during the preparation phase, ultimately leading to a higher quality teaching workforce.

Figure 8

Experimental Design and Study Protocol

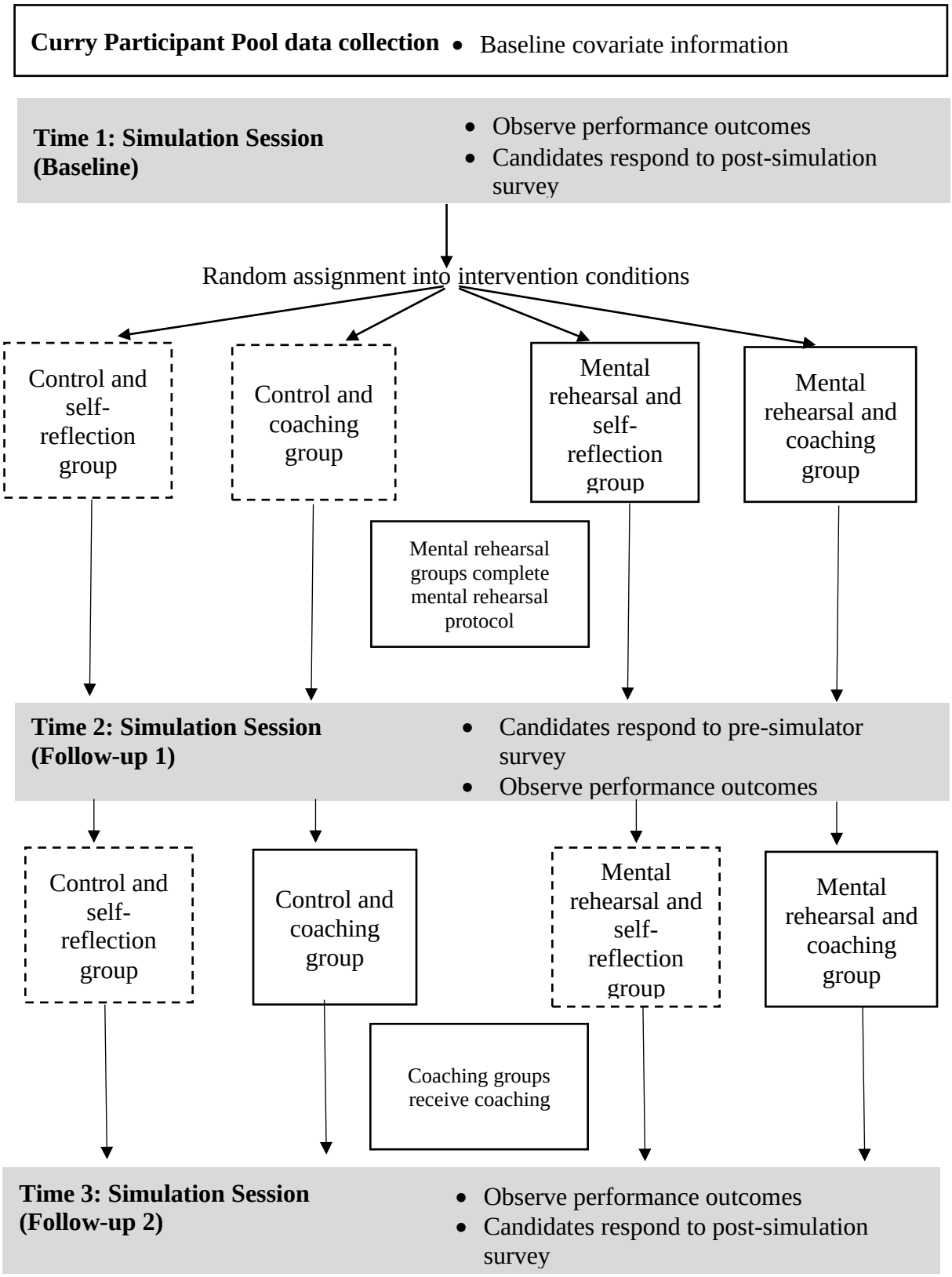


Figure 9

Theory Of Change for Mental Rehearsal Intervention

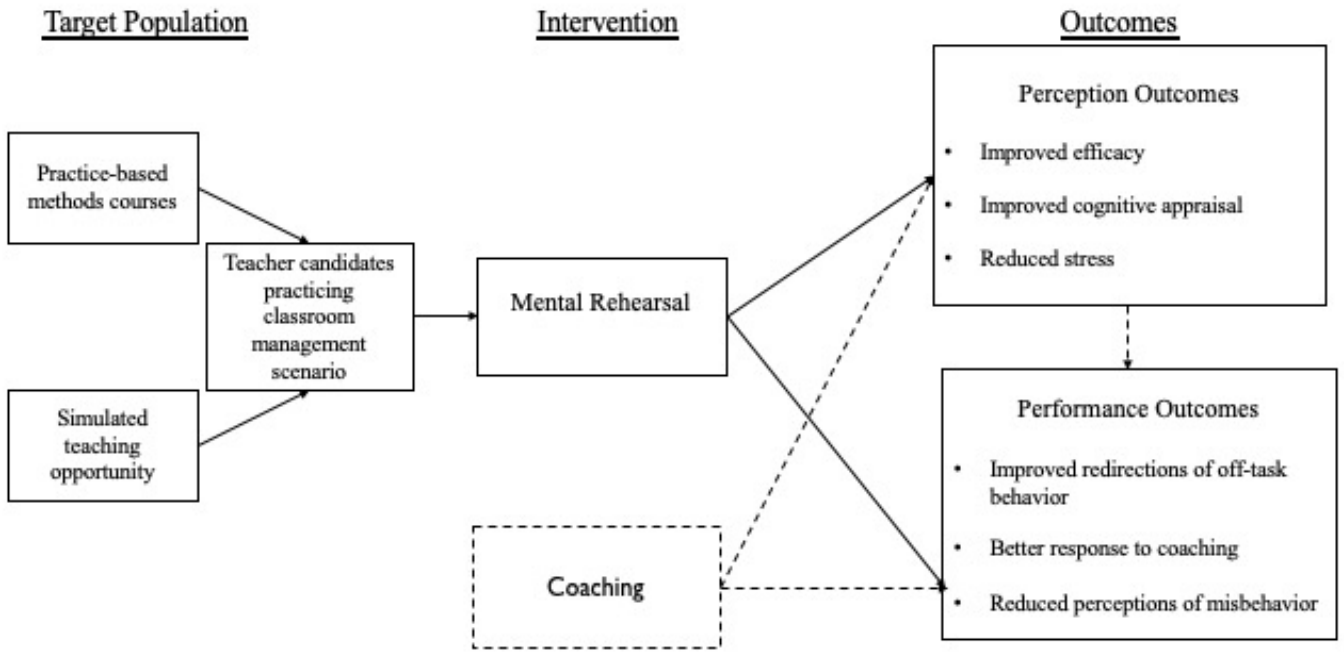


Figure 10

Improvements in Overall Quality Score by Mental Rehearsal Condition

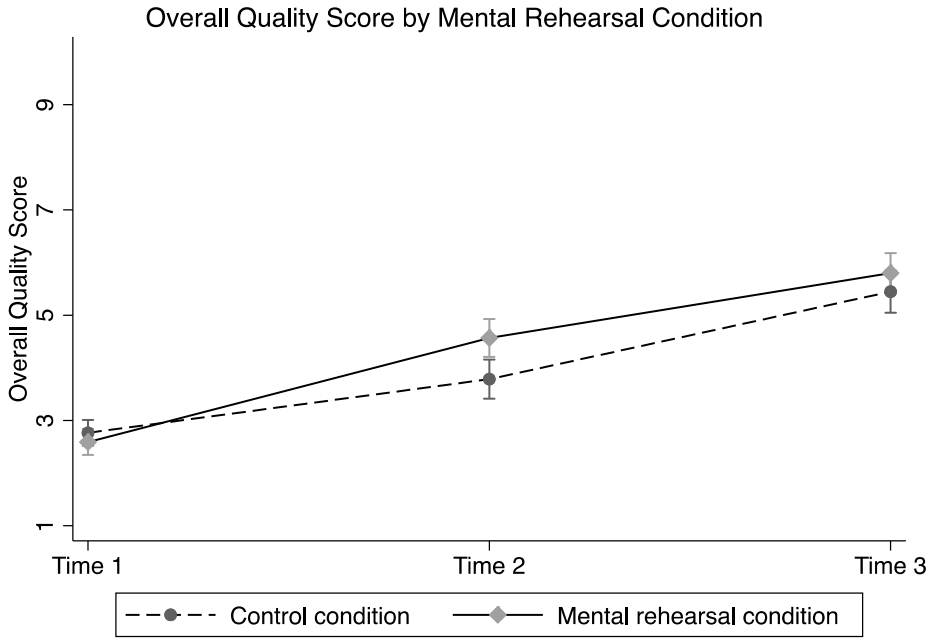


Figure 11

Improvements in Self-Efficacy Beliefs in Classroom Management by Mental Rehearsal Condition

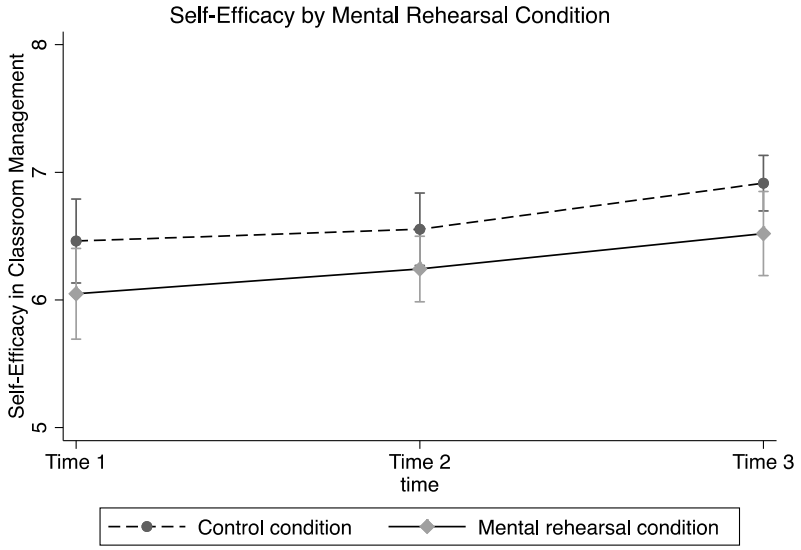


Figure 12

Improvements in Overall Quality Score by Mental Rehearsal and Coaching Conditions

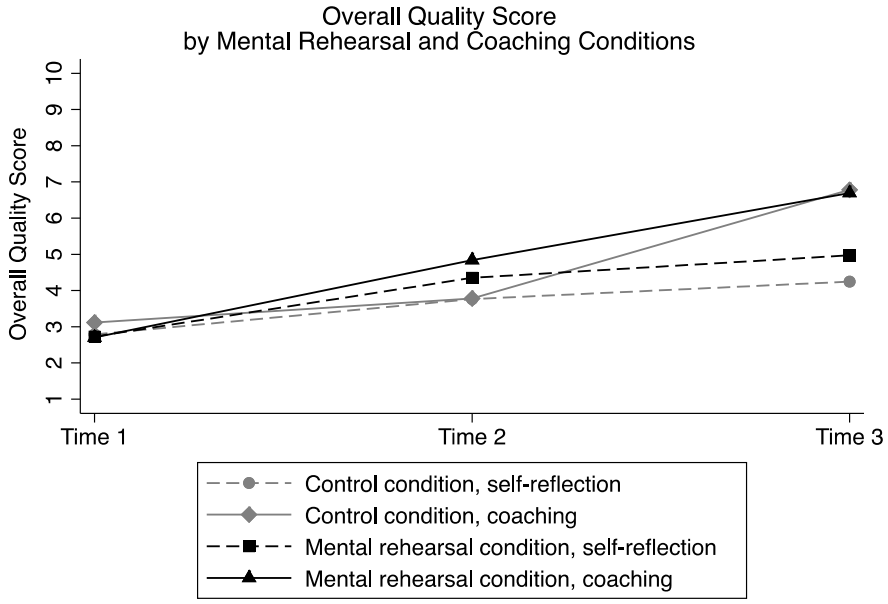


Figure 13

Improvements in Self-Efficacy Beliefs in Classroom Management by Mental Rehearsal and Coaching Conditions

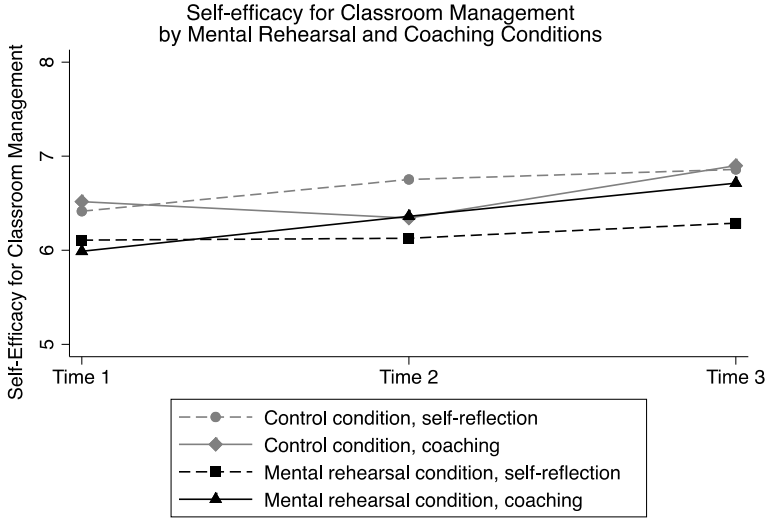


Table 11

Descriptive Statistics by Mental Rehearsal Condition

Pre-treatment Covariate	Control group Mean	Mental Rehearsal group Mean	Raw Difference	Glass's Delta	*p*	N
GPA	3.37	3.39	0.02	0.01	0.94	112
% Either parent a teacher	0.36	0.35	-0.01	-0.02	0.94	112
% Mother education- college or above	0.48	0.56	0.07	0.18	0.24	112
% Father education- college or above	0.68	0.69	0.00	0.04	0.96	112
% Female	1.05	1.10	0.05	0.13	0.48	112
Average Age	20.64	20.01	-0.62	-0.08	0.60	112
% White	0.59	0.66	0.07	0.15	0.43	112
Location of high school attended						
% Rural	0.02	0.00	-0.02	-0.13	0.32	112
% Suburban	0.20	0.20	0.00	0.00	0.97	112
% Urban	0.76	0.79	0.02	0.07	0.73	112
Average SES of high school attended						
% Low SES	0.02	0.00	-0.02	-0.13	0.32	112
% Middle SES	0.09	0.08	-0.02	-0.13	0.33	112
% High SES	0.56	0.44	-0.12	-0.21	0.11	112
Majority race of high school attended						
% Primarily students of color	0.02	0.00	-0.02	-0.13	0.32	112
% Mixed	0.07	0.07	0.00	0.00	1.00	112
% Primarily white students	0.32	0.31	-0.01	0.00	0.92	112
Average achievement level of high school attended						

% Primarily low achieving	0.02	0.00	-0.02	-0.13	0.32	112
% Primarily middle achieving	0.49	0.49	0.00	0.00	0.95	112
% Primarily high achieving	0.49	0.51	0.02	0.07	0.75	112
Baseline performance outcomes						
Overall Quality Score	2.14	1.94	-0.20	-0.19	0.21	112
Timeliness of Redirections	16.95	15.37	-1.58	-0.19	0.25	112
Proportion of Specific Redirections	0.17	0.13	-0.04	-0.14	0.43	112
Succinctness of Redirections	28.62	30.54	1.93	0.33	0.08	112
Baseline survey outcomes (post-coaching)						
Self-Efficacy- Classroom Management	5.75	5.32	-0.43	-0.31	0.08	112
Behavior Ratings Scale	2.86	2.93	0.07	0.13	0.47	112
Punitive Management Approaches Scale	4.91	5.13	0.22	0.13	0.50	112

Note. Demographic information comes from surveys administered to the teacher candidates and undergraduate participants. Each row represents results from a separate regression with the same right-hand side specification but a different baseline covariate as the dependent variable. Models include controls for randomization blocks. Glass' Delta was calculated using the mean difference between the control group and mental rehearsal group, and then dividing the result by the standard deviation of the control group. I also conducted a multivariate regression model for all 21 dependent variables, predicted by treatment status: $F(21, 55) = 1.60$, Prob > F = 0.838. The only covariate that was significantly different between the mental rehearsal group and the control group is self-efficacy beliefs at baseline. SES = socio-economic status. $^{+}p < .10$. $*p < .05$. $**p <$.

Table 12

Factor Analysis and Reliability for Construct Measures

Dimension	Number of factors identified using CFA	Range of item loadings	Chi-square and p-value	RMSEA	CFI	Reliability (alpha)
State Anxiety	1	0.55-1.22	$X^2\ (7) = 6.69, p > 0.10$	0.00	1.00	0.85
Oppositional/ Defiant Behavior Ratings	1	0.22-0.95	$X^2\ (3) = 13.52, p < 0.05$	0.128	0.975	0.83
Self-efficacy for classroom management	1	0.91-1.33	$X^2\ (13) = 19.25, p > 0.10$	0.04	0.99	0.94

Table 13

Longitudinal Results of Impact of Mental Rehearsal on Teacher Candidates' Outcomes

	Adjusted Means at Baseline (1)	Adjusted Means at Follow-up 1 (after Mental Rehearsal) (2)	Adjusted Means at Follow-up 1 (after Coaching) (3)
STAI			
Mental Rehearsal Group	-	11.79	-
	-	(19.18)	-
Control Group	-	11.27	-
	-	(19.22)	-
Difference between Mental Rehearsal and Control Groups	-	0.53	-
	-	(0.69)	-

Overall Quality Score			
Mental Rehearsal Group	2.70	4.58	5.76
	(0.12)	(0.18)	(0.19)
Control Group	2.95	3.78	5.44
	(0.12)	(0.19)	(2.70)
Difference between Mental Rehearsal and Control Groups	-0.25	0.81**	0.32
	(0.18)	(0.26)	(0.27)
Timeliness of Redirections			
Mental Rehearsal Group	15.10	10.78	7.07
	(0.83)	(0.61)	(0.42)
Control Group	15.18	14.98	9.09
	(0.84)	(0.68)	(0.56)
Difference between Mental Rehearsal and Control Groups	-0.09	-4.20**	-2.01**
	(1.18)	(0.93)	(0.71)
Proportion of Specific Redirections			
Mental Rehearsal Group	0.19	0.34	0.54
	(0.03)	(0.04)	(0.04)
Control Group	0.19	0.16	0.47
	(0.04)	(0.03)	(0.04)
Difference between Mental Rehearsal and Control Groups	0.00	0.18**	0.06
	(0.05)	(0.05)	(0.06)
Succinctness of Redirections			
Mental Rehearsal Group	30.78	25.09	16.73
	(0.93)	(1.18)	(1.14)
Control Group	27.68	27.70	19.57
	(0.98)	(1.04)	(1.31)
Difference between Mental Rehearsal and Control Groups	3.09*	-2.62+	-2.83
	(1.37)	(1.57)	(1.73)
Self-Efficacy for Classroom Management			
Mental Rehearsal Group	6.07	6.24	6.49
	(0.18)	(0.13)	(0.16)
Control Group	6.46	6.55	6.88
	(0.18)	(0.14)	(0.11)
Difference between Mental Rehearsal and Control Groups	-0.39	-0.30	-0.39+
	(0.26)	(0.20)	(0.20)
Oppositional/Defiant Behavior Ratings			

Mental Rehearsal Group	2.48	-	2.25
	(0.29)	-	(0.29)
Control Group	2.66	-	2.73
	(0.29)	-	(0.30)
Difference between Mental Rehearsal and Control Groups	-0.19	-	-0.48*
	(0.20)	-	(0.20)
Punitive Management Approach Endorsements			
Mental Rehearsal Group	4.71	-	4.49
	(0.85)	-	(0.79)
Control Group	5.09	-	4.79
	(0.84)	-	(0.81)
Difference between Mental Rehearsal and Control Groups	-0.38	-	-0.30
	(0.52)	-	(0.53)
Randomization blocks	x	x	x
Covariates	x	x	x
Coaching condition	x	x	x

Note. Adjusted control group means and clustered standard errors (in parentheses) are reported in Column (1). Coefficients and clustered standard errors are reported in Columns (2) through (4) represent mean adjusted differences between Time 1 and Time 2, Time 2 and Time 3 and Time 3 and Time 1 for control and mental rehearsal conditions. Coefficients are taken from separate regressions of the outcome on intervention assignment and time variable. All models include controls for randomization blocks, whether teacher candidate was above the age of 21 at baseline and coaching condition assignment. I include missing data indicators in cases where respondents have missing pre-treatment covariates. Glass' Delta was calculated using the mean difference between the control group and mental rehearsal group, and then dividing the result by the standard deviation of the control group †p < .10. *p < .05. **p < .01.

Table 14

Longitudinal Results of Impact of Mental Rehearsal and Coaching on Teacher Candidates' Outcomes

	Adjusted Means at Baseline (1)	Adjusted Means at Follow-up 1 (after Mental Rehearsal) (2)	Adjusted Means at Follow-up 1 (after Coaching) (3)
STAI			
Mental rehearsal condition, coaching group	16.71	-	-
	(2.13)	-	-
Control condition, coaching group	17.21	-	-
	(2.33)	-	-
Difference between Mental Rehearsal and Control Groups	-0.49	-	-
	(1.00)	-	-
Mental rehearsal condition, self-reflection group	16.28	-	-
	(2.29)	-	-
Control condition, self-reflection group	14.77	-	-
	(2.42)	-	-
Difference between Mental Rehearsal and Control Groups	1.50	-	-
	(0.97)	-	-
Overall Quality Score			
Mental rehearsal condition, coaching group	2.88	5.02	6.89
	(0.44)	(0.48)	(0.49)
Control condition, coaching group	3.24	3.92	6.93
	(0.41)	(0.46)	(0.45)
Difference between Mental Rehearsal and Control Groups	-0.36	1.10**	-0.04
	(0.24)	(0.36)	(0.34)
Mental rehearsal condition, self-reflection group	2.88	4.51	5.13
	(0.44)	(0.49)	(0.46)
Control condition, self-reflection group	2.97	3.94	4.42
	(0.40)	(0.46)	(0.45)
Difference between Mental Rehearsal and Control Groups	-0.08	0.56	0.70*
	(0.22)	(0.38)	(0.34)
Timeliness of Redirections			
Mental rehearsal condition, coaching group	12.69	8.80	3.77
	(2.17)	(1.97)	(1.83)

Control condition, coaching group	11.78	12.21	3.84
	(2.06)	(2.11)	(1.80)
Difference between Mental Rehearsal and Control Groups	0.91	-3.41*	-0.07
	(1.57)	(1.36)	(0.74)
Mental rehearsal condition, self-reflection group	13.48	8.57	6.09
	(2.18)	(2.03)	(1.95)
Control condition, self-reflection group	14.31	13.40	9.72
	(2.62)	(2.02)	(1.97)
Difference between Mental Rehearsal and Control Groups	-0.83	-4.82**	-3.63**
	(1.77)	(1.28)	(1.13)
Proportion of Specific Redirections			
Mental rehearsal condition, coaching group	0.23	0.41	0.74
	(0.11)	(0.13)	(0.12)
Control condition, coaching group	0.27	0.19	0.73
	(0.11)	(0.11)	(0.12)
Difference between Mental Rehearsal and Control Groups	-0.04	0.21*	0.01
	(0.07)	(0.07)	(0.08)
Mental rehearsal condition, self-reflection group	0.21	0.34	0.41
	(0.12)	(0.13)	(0.12)
Control condition, self-reflection group	0.23	0.22	0.33
	(0.12)	(0.11)	(0.12)
Difference between Mental Rehearsal and Control Groups	-0.02	0.12+	0.08
	(0.06)	(0.07)	(0.09)
Succinctness of Redirections			
Mental rehearsal condition, coaching group	28.79	21.02	9.18
	(2.99)	(3.15)	(3.21)
Control condition, coaching group	25.40	26.68	9.16
	(2.66)	(3.00)	(2.62)
Difference between Mental Rehearsal and Control Groups	3.39*	-5.66*	0.02
	(1.66)	(2.48)	(1.76)
Mental rehearsal condition, self-reflection group	28.29	24.53	19.18
	(2.68)	(3.11)	(3.00)
Control condition, self-reflection group	25.91	24.75	24.87
	(2.92)	(2.91)	(3.41)
Difference between Mental Rehearsal and Control Groups	2.38	-0.22	-5.69*
	(1.80)	(1.98)	(2.53)
Self-Efficacy for Classroom Management			
Mental rehearsal condition, coaching group	5.52	5.90	6.25
	(0.58)	(0.53)	(0.55)
Control condition, coaching group	5.96	5.78	6.36
	(0.54)	(0.50)	(0.53)

Difference between Mental Rehearsal and Control Groups	-0.44 (0.37)	0.12 (0.26)	-0.11 (0.28)
Mental rehearsal condition, self-reflection group	5.54 (0.53)	5.57 (0.53)	5.73 (0.53)
Control condition, self-reflection group	5.93 (0.57)	6.26 (0.58)	6.37 (0.54)
Difference between Mental Rehearsal and Control Groups	-0.38 (0.33)	-0.70* (0.28)	-0.64* (0.28)
Oppositional/ Defiant Behavior Ratings			
Mental rehearsal condition, coaching group	2.98 (0.31)	- -	2.31 (0.29)
Control condition, coaching group	2.61 (0.24)	- -	2.32 (0.25)
Difference between Mental Rehearsal and Control Groups	0.37+ (0.21)	- -	-0.02 (0.15)
Mental rehearsal condition, self-reflection group	2.48 (0.29)	- -	2.25 (0.29)
Control condition, self-reflection group	2.66 (0.29)	- -	2.73 (0.30)
Difference between Mental Rehearsal and Control Groups	-0.19 (0.20)	- -	-0.48* (0.20)
Punitive Management Approach Endorsements			
Mental rehearsal condition, coaching group	4.91 (0.82)	- -	4.31 (0.80)
Control condition, coaching group	4.36 (0.78)	- -	4.36 (0.82)
Difference between Mental Rehearsal and Control Groups	0.55 (0.45)	- -	-0.05 (0.51)
Mental rehearsal condition, self-reflection group	4.71 (0.85)	- -	4.49 (0.79)
Control condition, self-reflection group	5.09 (0.84)	- -	4.79 (0.81)
Difference between Mental Rehearsal and Control Groups	-0.38 (0.52)	- -	-0.30 (0.53)
Randomization blocks	x	x	x
Covariates	x	x	x
Coaching condition	x	x	x

Note. Adjusted control group means and clustered standard errors (in parentheses) are reported in Column (1). Coefficients and clustered standard errors are reported in Columns (2) through (4) represent mean adjusted differences between Time 1 and Time 2, Time 2 and Time 3 and Time 3 and Time 1 for control and mental rehearsal conditions. Coefficients are taken from separate regressions of the outcome on intervention assignment and time variable. All models include controls for randomization blocks, whether teacher candidate was above the age of 21 at baseline. I include missing data indicators in cases where respondents have missing pre-treatment covariates. Glass' Delta was calculated using the mean difference between the control group and mental rehearsal group, and then dividing the result by the standard deviation of the control group †p < .10. *p < .05. **p < .01.

REFERENCES

Alt, D. (2015). Assessing the contribution of a constructivist learning environment to academic self-efficacy in higher education. *Learning Environments Research, 18*(1), 47–67. https://doi.org/10.1007/s10984-015-9174-5

American Association of Colleges for Teacher Education. (2013). The changing teacher preparation profession.

Anderson, R., Greene, M., & Loewen, P. (1988). Relationships among teachers' and students' thinking skills, sense of efficacy, and student achievement. Alberta Journal of Educational Research, 34(2), 148-165

Anderson, N. A., & Radencich, M. C. (2001). The Value of Feedback in an Early Field Experience: Peer, Teacher, and Supervisor Coaching. *Action in Teacher Education*, *23*(3), 66–74. https://doi.org/10.1080/01626620.2001.10463076

Anglin, K. L., & Wong, V. C. (2020). Using Semantic Similarity to Assess Adherence and Replicability of Intervention Delivery.

Armor, D., Conroy-Oseguera, P., Cox M., King, N., McDonnell, L., Pascal, A. Pauly, E., & Zellman, G. (1976). Analysis of the school preferred reading programs in selected Los Angeles minority schools. (REPORT NO. R-2007-LAUSD). Santa Monica, CA: Rand Corporation. (ERIC Document Reproduction Service No. 130 243).

Arora, S., Aggarwal, R., Sirimanna, P., Moran, A., Grantcharov, T., Kneebone, R., Sevdalis, N., & Darzi, A. (2011). Mental Practice Enhances Surgical Technical Skills. *Annals of Surgery*, *253*(2), 265–270. https://doi.org/10.1097/sla.0b013e318207a789

Arora, S., Miskovic, D., Hull, L., Moorthy, K., Aggarwal, R., Johannsson, H., Gautama, S., Kneebone, R., & Sevdalis, N. (2011). Self vs expert assessment of technical and non-technical skills in high fidelity simulation. *The American Journal of Surgery*, *202*(4), 500–506. https://doi.org/10.1016/j.amjsurg.2011.01.024

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do First Impressions Matter? Predicting Early Career Teacher Effectiveness. *AERA Open*, *1*(4), 233285841560783. https://doi.org/10.1177/2332858415607834

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295x.84.2.191

Bandura, A. (1986). The Explanatory and Predictive Scope of Self-Efficacy Theory. *Journal of Social and Clinical Psychology*, *4*(3), 359–373. https://doi.org/10.1521/jscp.1986.4.3.359

Bandura, A. (1997a). The nature and structure of self-efficacy. *Self-efficacy: the exercise of control. New York, NY: WH Freeman and Company*, 37-78.

Bandura, A. (1997b). Self-efficacy and health behaviour. *Cambridge handbook of psychology, health and medicine*, 160-162.

Bautista, N. U., & Boone, W. J. (2015). Exploring the Impact of TeachME™ Lab Virtual Classroom Teaching Simulation on Early Childhood Education Majors' Self-Efficacy Beliefs. *Journal of Science Teacher Education*, *26*(3), 237–262. https://doi.org/10.1007/s10972-014-9418-8

Beijaard, D., Korthagen, F., & Verloop, N. (2007). Understanding how teachers learn as a prerequisite for promoting teacher learning. *Teachers and Teaching*, *13*(2), 105–108. https://doi.org/10.1080/13540600601152298

Bell, C. A., Jones, N. D., Qi, Y., & Lewis, J. M. (2018). Strategies for Assessing Classroom Teaching: Examining Administrator Thinking as Validity Evidence. *Educational Assessment*, *23*(4), 229–249. https://doi.org/10.1080/10627197.2018.1513788

Beran, T. (2005). A new perspective on managing school bullying: Pre-service teachers' attitudes. *Journal of Social Sciences*, *8*(3), 43-49.

Blascovich, J., Seery, M. D., Mugridge, C. A., Norris, R., & Weisbuch, M. (2004). Predicting athletic performance from cardiovascular indexes of challenge and threat. *Journal of Experimental Social Psychology*, *40*(5), 683–688. https://doi.org/10.1016/j.jesp.2003.10.007

Blazar, D., & Kraft, M. A. (2015). Exploring Mechanisms of Effective Teacher Coaching. *Educational Evaluation and Policy Analysis*, *37*(4), 542–566. https://doi.org/10.3102/0162373715579487

Bowman, C. L., & McCormick, S. (2000). Comparison of Peer Coaching Versus Traditional Supervision Effects. *The Journal of Educational Research*, *93*(4), 256–261. https://doi.org/10.1080/00220670009598714

Braun, H. (2015). The Value in Value Added Depends on the Ecology. *Educational Researcher*, *44*(2), 127–131. https://doi.org/10.3102/0013189x15576341

Britton, L. R., & Anderson, K. A. (2010). Peer coaching and pre-service teachers: Examining an underutilised concept. *Teaching and Teacher Education*, *26*(2), 306–314. https://doi.org/10.1016/j.tate.2009.03.008

Brown, A. L., Lee, J., & Collins, D. (2014). Does student teaching matter? Investigating pre-service teachers' sense of efficacy and preparedness. *Teaching Education*, *26*(1), 77–93. https://doi.org/10.1080/10476210.2014.957666

Brunsting, N. C., Sreckovic, M. A., & Lane, K. L. (2014). Special Education Teacher Burnout: A Synthesis of Research from 1979 to 2013. *Education and Treatment of Children*, *37*(4), 681–711. https://doi.org/10.1353/etc.2014.0032

Busch, D. (2010). Pre-service teacher beliefs about language learning: The second language acquisition course as an agent for change. *Language Teaching Research*, *14*(3), 318–337. https://doi.org/10.1177/1362168810365239

Cantrell, P., Young, S., & Moore, A. (2003). Factors Affecting Science Teaching Efficacy of Preservice Elementary Teachers. *Journal of Science Teacher Education*, *14*(3), 177–192. https://doi.org/10.1023/a:1025974417256

Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement: A study at the school level. *Journal of School Psychology*, *44*(6), 473–490. https://doi.org/10.1016/j.jsp.2006.09.001

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*(5), 619–636. https://doi.org/10.1002/acp.1101

Carroll, D. (2012). Examining the Development of Dispositions for Ambitious Teaching: One Teacher Candidate's Journey. *The New Educator*, *8*(1), 38–64. https://doi.org/10.1080/1547688x.2011.619950

Chang, M. L. (2009). An Appraisal Perspective of Teacher Burnout: Examining the Emotional Work of Teachers. *Educational Psychology Review*, *21*(3), 193–218. https://doi.org/10.1007/s10648-009-9106-y

Chaplain, R. P. (2008). Stress and psychological distress among trainee secondary teachers in England. *Educational Psychology*, *28*(2), 195–209. https://doi.org/10.1080/01443410701491858

Clark, S., & Newberry, M. (2018). Are we building preservice Teacher self-efficacy? A large-scale study examining Teacher education experiences. *Asia-Pacific Journal of Teacher Education*, *47*(1), 32–47. https://doi.org/10.1080/1359866x.2018.1497772

Coburn, C. E., & Woulfin, S. L. (2012). Reading Coaches and the Relationship Between Policy and Practice. *Reading Research Quarterly*, *47*(1), 5–30. https://doi.org/10.1002/rrq.008

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher Coaching in a Simulated Environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231. https://doi.org/10.3102/0162373720906217

Cohen, J., & Wiseman, E. A. (2021). Teacher Supports for Text-Based Instruction: Experimental Evidence from a Mixed-Reality Simulation in Preservice Preparation.

Cole, S. R., & Stuart, E. A. (2010). Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *American Journal of Epidemiology*, *172*(1), 107–115. https://doi.org/10.1093/aje/kwq084

Corcoran, E. (1981). Transition Shock. *Journal of Teacher Education*, *32*(3), 19–23. https://doi.org/10.1177/002248718103200304

Costa, P. T., Jr., & McCrae, R. R. (1992). Revised NEO Personality Inventory (NEO–PI–R) and NEO Five-Factor Inventory (NEO–FFI) professional manual. Odessa, FL: Psychological Assessment Resources

Cruickshank, D. R., & Armaline, W. D. (1986). Field Experiences in Teacher Education: Considerations and Recommendations. *Journal of Teacher Education*, *37*(3), 34–40. https://doi.org/10.1177/002248718603700307

Da Silva, M. (2005). Constructing the Teaching Process from Inside Out: How Pre-Service Teachers Make Sense of Their Perceptions of the Teaching of the Four Skills. *TESL-EJ*, *9*(2), n2.

Damon, W. (2007). Dispositions and Teacher Assessment. *Journal of Teacher Education*, *58*(5), 365–369. https://doi.org/10.1177/0022487107308732

Darling-Hammond, L. (2000). How Teacher Education Matters. *Journal of Teacher Education*, *51*(3), 166–173. https://doi.org/10.1177/0022487100051003002

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Vasquez Heilig, J. (2005). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness. *Education Policy Analysis Archives*, *13*, 42. https://doi.org/10.14507/epaa.v13n42.2005

Darling-Hammond, L. (2015). Can Value Added Add Value to Teacher Evaluation? *Educational Researcher*, *44*(2), 132–137. https://doi.org/10.3102/0013189x15575346

de Lange, F. P., Roelofs, K., & Toni, I. (2008). Motor imagery: A window into the mechanisms and alterations of the motor system. *Cortex*, *44*(5), 494–506. https://doi.org/10.1016/j.cortex.2007.09.002

Decker, L. E., & Rimm-Kaufman, S. E. (2008). Personality characteristics and teacher beliefs among pre-service teachers. *Teacher Education Quarterly*, *35*(2), 45-64.

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher*, *38*(3), 181–199. https://doi.org/10.3102/0013189x08331140

Desimone, L. M., & Pak, K. (2016). Instructional Coaching as High-Quality Professional Development. *Theory Into Practice*, *56*(1), 3–12. https://doi.org/10.1080/00405841.2016.1241947

Deussen, T., Coskie, T., Robinson, L., & Autio, E. (2007). "Coach" can mean many things: Five categories of literacy coaches in Reading First. *Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education*.

Dial, E. (2015). An experienced practical reflection: Creating confident, effective teachers by changing pre-service programs to promote resilience, assessment literacy, and collaboration. National Teacher Education Journal, 8(1). Retrieved from http://www.ntejournal.com/

Dicke, T., Parker, P. D., Marsh, H. W., Kunter, M., Schmeck, A., & Leutner, D. (2014). Self-efficacy in classroom management, classroom disturbances, and emotional exhaustion: A moderated mediation analysis of teacher candidates. *Journal of Educational Psychology*, *106*(2), 569–583. https://doi.org/10.1037/a0035504

Dieker, L. A., Rodriguez, J. A., Lignugaris/Kraft, B., Hynes, M. C., & Hughes, C. E. (2013). The Potential of Simulated Environments in Teacher Education. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *37*(1), 21–33. https://doi.org/10.1177/0888406413512683

Djigić, G., Stojiljković, S., & Dosković, M. (2014). Basic Personality Dimensions and Teachers' Self-efficacy. *Procedia - Social and Behavioral Sciences*, *112*, 593–602. https://doi.org/10.1016/j.sbspro.2014.01.1206

Driscoll, K. C., & Pianta, R. C. (2010). Banking Time in Head Start: Early Efficacy of an Intervention Designed to Promote Supportive Teacher–Child Relationships. *Early Education & Development*, *21*(1), 38–64. https://doi.org/10.1080/10409280802657449

Driskell, J. E., Copper, C., & Moran, A. (1994). Does mental practice enhance performance? *Journal of Applied Psychology*, *79*(4), 481–492. https://doi.org/10.1037/0021-9010.79.4.481

Druckman, D., & Swets, J. A. (Eds.). (1988). *Enhancing human performance: Issues, theories, and techniques*. National Academies Press.

Eldred-Evans, D., Grange, P., Cheang, A., Yamamoto, H., Ayis, S., Mulla, M., Immenroth, M., Sharma, D., & Reedy, G. (2013). Using the Mind as a Simulator: A Randomized Controlled Trial of Mental Training. *Journal of Surgical Education*, *70*(4), 544–551. https://doi.org/10.1016/j.jsurg.2013.04.003

Elmore, R. F. (2006). Foreword. In Fullan, M., Hill, P., and Crévola, C., Breakthrough (pp. xixiv). Th ousand Oaks, CA: Corwin Press.

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the new science of expertise*. Houghton Mifflin Harcourt.

Ethel, R. G., & McMeniman, M. M. (2000). Unlocking the Knowledge in Action of an Expert Practitioner. *Journal of Teacher Education*, *51*(2), 87–101. https://doi.org/10.1177/002248710005100203

Evans, E. D., & Tribble, M. (1986). Perceived Teaching Problems, Self-Efficacy, and Commitment to Teaching among Preservice Teachers. *The Journal of Educational Research*, *80*(2), 81–85. https://doi.org/10.1080/00220671.1986.10885728

Fang, Z. (1996). A review of research on teacher beliefs and practices. *Educational Research*, *38*(1), 47–65. https://doi.org/10.1080/0013188960380104

Feiman-Nemser, S. (2003). What new teachers need to learn. *Educational leadership*, *60*(8), 25-29.

Feldman, P. J., Cohen, S., Hamrick, N., & Lepore, S. J. (2004). Psychological stress, appraisal, emotion and Cardiovascular response in a public speaking task. *Psychology & Health*, *19*(3), 353–368. https://doi.org/10.1080/0887044042000193497

Feltz, D. L., Landers, D. M., & Becker, B. J. (1988). *A revised meta-analysis of the mental practice literature on motor skill learning*.

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. https://doi.org/10.1016/j.tate.2006.03.013

Fletcher, S., & Mullen, C. A. (Eds.). (2012). *Sage handbook of mentoring and coaching in education*. Sage.

Flower, A., McKenna, J. W., & Haring, C. D. (2016). Behavior and classroom management: Are teacher preparation programs really preparing our teachers? *Preventing School Failure: Alternative Education for Children and Youth*, *61*(2), 163–169. https://doi.org/10.1080/1045988x.2016.1231109

Friedman, Y., & Gavish, B. (2001). The beginner teacher: Difficulties, support and adjustment. *Jerusalem: Henrietta Szold institute*.

Friedman, I. A. (2006). Classroom Management and Teacher Stress and Burnout.

Freytag, C. E. (2001). Teacher Efficacy and Inclusion: The Impact of Pre-service Experiences on Beliefs.

Gavish, B., & Friedman, I. A. (2010). Novice teachers' experience of teaching: a dynamic aspect of burnout. *Social Psychology of Education*, *13*(2), 141–167. https://doi.org/10.1007/s11218-009-9108-0

George, S. V., Richardson, P. W., & Watt, H. M. (2018). Early career teachers' self-efficacy: A longitudinal study from Australia. *Australian Journal of Education*, *62*(2), 217–233. https://doi.org/10.1177/0004944118779601

Goldhaber, D. (2015). Exploring the Potential of Value-Added Performance Measures to Affect the Quality of the Teacher Workforce. *Educational Researcher*, *44*(2), 87–95. https://doi.org/10.3102/0013189x15574905

Goldhaber, D. (2018). Evidence-Based Teacher Preparation: Policy Context and What We Know. *Journal of Teacher Education*, *70*(2), 90–101. https://doi.org/10.1177/0022487118800712

Good, T. L., & Brophy, J. E. (2003). Looking in classrooms (9e éd.). *Boston, MA: Allyn et Bacon*.

Gordon, L. M. (2001). High Teacher Efficacy as a Marker of Teacher Effectiveness in the Domain of Classroom Management.

Greenberg, J., & Jacobs, S. (2009). Preparing Tomorrow's Teachers: Are Colorado's Education School Graduates Ready to Teach Reading and Mathematics in Elementary Classrooms?. *National Council on Teacher Quality*.

Grossman, P. L., Wilson, S. M., & Shulman, L. S. (1989). Teachers of substance: Subject matter knowledge for teaching. *Profesorado, Revista de currículum y formación del profesorado*, *9*(2), 1-25.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009a). Teaching practice: A cross-professional perspective. *Teachers college record*, *111*(9), 2055-2100.

Grossman, P., Hammerness, K., & McDonald, M. (2009b). Redefining teaching, re-imagining teacher education. *Teachers and Teaching*, *15*(2), 273–289. https://doi.org/10.1080/13540600902875340

Guiney, E. (2001). Coaching isn't just for athletes: The role of teacher leaders. *Phi Delta Kappan*, *82*(10), 740-743.

Gurvitch, R., & Metzler, M. W. (2009). The effects of laboratory-based and field-based practicum experience on pre-service teachers' self-efficacy. *Teaching and Teacher Education*, *25*(3), 437–443. https://doi.org/10.1016/j.tate.2008.08.006

Hammond, L., & Moore, W. (2018). Teachers Taking up Explicit Instruction: The Impact of a Professional Development and Directive Instructional Coaching Model. *Australian Journal of Teacher Education*, *43*(7), 110–133. https://doi.org/10.14221/ajte.2018v43n7.7

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7–8), 798–812. https://doi.org/10.1016/j.jpubeco.2010.11.009

Harvey, A., Nathens, A. B., Bandiera, G., & LeBlanc, V. R. (2010). Threat and challenge: cognitive appraisal and stress responses in simulated trauma resuscitations. *Medical Education*, *44*(6), 587–594. https://doi.org/10.1111/j.1365-2923.2010.03634.x

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Erratum: Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(2), 164–165. https://doi.org/10.1002/jrsm.17

Helgevold, N., Næsheim-Bjørkvik, G., & Østrem, S. (2015). Key focus areas and use of tools in mentoring conversations during internship in initial teacher education. *Teaching and Teacher Education*, *49*, 128–137. https://doi.org/10.1016/j.tate.2015.03.005

Higgins, J. P. T. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional Development Research. *Educational Researcher*, *42*(9), 476–487. https://doi.org/10.3102/0013189x13512674

Hoffman, J. V., Wetzel, M. M., Maloch, B., Greeter, E., Taylor, L., DeJulio, S., & Vlach, S. K. (2015). What can we learn from studying the coaching interactions between cooperating teachers and preservice teachers? A literature review. *Teaching and Teacher Education*, *52*, 99–112. https://doi.org/10.1016/j.tate.2015.09.004

Hong, J., Greene, B., Roberson, R., Cross Francis, D., & Rapacki Keenan, L. (2017). Variations in pre-service teachers' career exploration and commitment to

teaching. *Teacher Development*, *22*(3), 408–426. https://doi.org/10.1080/13664530.2017.1358661
Hopkins, W. S., Hoffman, S. Q., & Moss, V. D. (1997). Professional development schools and pre-service teacher stress. *Action in Teacher Education*, *18*(4), 36-46.
Howard, T. C., & Milner, H. R. (2014). Teacher preparation for urban schools. *Handbook of urban education*, *11*(3), 199-216.
Huberman, M. (1989). The professional life cycle of teachers. *Teachers college record*.
Humphrey, D. C., & Wechsler, M. E. (2006). Fighting the wrong battle in the teacher preparation wars. *Education Week*, *26*(1).
Ignacio, J., Scherpbier, A., Dolmans, D., Rethans, J. J., & Liaw, S. Y. (2017). Mental Rehearsal Strategy for Stress Management and Performance in Simulations. *Clinical Simulation in Nursing*, *13*(7), 295–302. https://doi.org/10.1016/j.ecns.2017.04.005
Ingersoll, R. M. (2001). Teacher Turnover and Teacher Shortages: An Organizational Analysis. *American Educational Research Journal*, *38*(3), 499–534. https://doi.org/10.3102/00028312038003499
Ippolito, J. (2010). Three Ways That Literacy Coaches Balance Responsive and Directive Relationships with Teachers. *The Elementary School Journal*, *111*(1), 164–190. https://doi.org/10.1086/653474
Israel, M., Knowlton, E., Griswold, D., & Rowland, A. (2009). Applications of Video-Conferencing Technology in Special Education Teacher Preparation. *Journal of Special Education Technology*, *24*(1), 15–25. https://doi.org/10.1177/016264340902400102
Jamil, F. M., Downer, J. T., & Pianta, R. C. (2012). Association of pre-service teachers' performance, personality, and beliefs with teacher self-efficacy at program completion. *Teacher Education Quarterly*, *39*(4), 119-138.
Jennings, P. A., & Greenberg, M. T. (2009). The Prosocial Classroom: Teacher Social and Emotional Competence in Relation to Student and Classroom Outcomes. *Review of Educational Research*, *79*(1), 491–525. https://doi.org/10.3102/0034654308325693
Jones, L., & Stuth, G. (1997). The uses of mental imagery in athletics: An overview. *Applied and Preventive Psychology*, *6*(2), 101–115. https://doi.org/10.1016/s0962-1849(05)80016-2
Kaff, M. S., Zabel, R. H., & Milham, M. (2007). Revisiting cost-benefit relationships of behavior management strategies: What special educators say about usefulness, intensity, and effectiveness. *Preventing School Failure: Alternative Education for Children and Youth*, *51*(2), 35-45.
Kagan, D. M. (1992). Implication of Research on Teacher Belief. *Educational Psychologist*, *27*(1), 65–90. https://doi.org/10.1207/s15326985ep2701_6

Kain, D. J. (2003). Teacher-Centered versus Student-Centered: Balancing Constraint and Theory in the Composition Classroom. *Pedagogy*, *3*(1), 104–108. https://doi.org/10.1215/15314200-3-1-104

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631. https://doi.org/10.1016/j.econedurev.2007.05.005

Kaner, S. (2010). Examining Teachers' Self-Efficacy Beliefs of Students With and Without Specıal Needs. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi*, *43*(1), 193-218.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Kavanagh, S. S., & Rainey, E. C. (2017). Learning to Support Adolescent Literacy: Teacher Educator Pedagogy and Novice Teacher Take Up in Secondary English Language Arts Teacher Preparation. *American Educational Research Journal*, *54*(5), 904–937. https://doi.org/10.3102/0002831217710423

Killion, J. (2016). Changes in coaching study design shed light on how features impact teacher practice. *The Learning Professional*, *37*(2), 58.

Kline, R. B. (2011). Convergence of structural equation modeling and multilevel modeling.

Kohler, F. W., Ezell, H. K., & Paluselli, M. (1999). Promoting Changes in Teachers' Conduct of Student Pair Activities. *The Journal of Special Education*, *33*(3), 154–165. https://doi.org/10.1177/002246699903300303

Kokkinos, C. M., Panayiotou, G., & Davazoglou, A. M. (2004). Correlates of teacher appraisals of student behaviors. *Psychology in the Schools*, *42*(1), 79–89. https://doi.org/10.1002/pits.20031

Kraft, M. A., & Papay, J. P. (2014). Can Professional Environments in Schools Promote Teacher Development? Explaining Heterogeneity in Returns to Teaching Experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500. https://doi.org/10.3102/0162373713519496

Kraft, M. A., & Blazar, D. (2016). Individualized Coaching to Improve Teacher Practice Across Grades and Subjects: New Experimental Evidence. *Educational Policy*, *31*(7), 1033–1068. https://doi.org/10.1177/0895904816631099

Kraft, M. A., Blazar, 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*(4), 547–588. https://doi.org/10.3102/0034654318759268

Kretlow, A. G., & Bartholomew, C. C. (2010). Using Coaching to Improve the Fidelity of Evidence-Based Practices: A Review of Studies. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for*

Exceptional Children, *33*(4), 279–299. https://doi.org/10.1177/0888406410371643
Krippendorff, K. (1970). Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology*, *2*, 139. https://doi.org/10.2307/270787
LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, 604-620.
Lampert, M. (2009). Learning Teaching in, from, and for Practice: What Do We Mean? *Journal of Teacher Education*, *61*(1–2), 21–34. https://doi.org/10.1177/0022487109347321
Latham, N. I., & Vogt, W. P. (2007). Do Professional Development Schools Reduce Teacher Attrition? *Journal of Teacher Education*, *58*(2), 153–167. https://doi.org/10.1177/0022487106297840
Lauermann, F., & Berger, J. L. (2021). Linking teacher self-efficacy and responsibility with teachers' self-reported and student-reported motivating styles and student engagement. *Learning and Instruction*, 101441. https://doi.org/10.1016/j.learninstruc.2020.101441
Lazarides, R., Watt, H. M., & Richardson, P. W. (2020). Teachers' classroom management self-efficacy, perceived classroom management and teaching contexts from beginning until mid-career. *Learning and Instruction*, *69*, 101346. https://doi.org/10.1016/j.learninstruc.2020.101346
Lazarus, R. S. (1966). Psychological stress and the coping process.
Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer publishing company.
Lazarus, R. S. (2006). *Stress and emotion: A new synthesis*. Springer Publishing Company.
LeBlanc, V. R. (2009). The effects of acute stress on performance: implications for health professions education. *Academic Medicine*, *84*(10), S25-S33.
Le Maistre, C., & Paré, A. (2010). Whatever it takes: How beginning teachers learn to survive. *Teaching and Teacher Education*, *26*(3), 559–564. https://doi.org/10.1016/j.tate.2009.06.016
Lee, C. (1990). Psyching Up for a Muscular Endurance Task: Effects of Image Content on Performance and Mood State. *Journal of Sport and Exercise Psychology*, *12*(1), 66–73. https://doi.org/10.1123/jsep.12.1.66
Leinhardt, G., Young, K. M., & Merriman, J. (1995). Integrating professional knowledge: The theory of practice and the practice of theory. *Learning and Instruction*, *5*(4), 401–408. https://doi.org/10.1016/0959-4752(95)00025-9
Leung, S. S., Chiang, V. C., Chui, Y., Mak, Y., & Wong, D. F. (2011). A Brief Cognitive-behavioral Stress Management Program for Secondary School Teachers. *Journal of Occupational Health*, *53*(1), 23–35. https://doi.org/10.1539/joh.l10037

Levine, A. (2006). Will universities maintain control of teacher education?. *Change: The Magazine of Higher Learning*, *38*(4), 36-43

Li, L. (2020). The Impact of Teacher Education on Teacher Cognition. In *Language Teacher Cognition* (pp. 203-239). Palgrave Macmillan, London

Lin, H. L., & Gorrell, J. (2001). Exploratory analysis of pre-service teacher efficacy in Taiwan. *Teaching and Teacher Education*, *17*(5), 623–635. https://doi.org/10.1016/s0742-051x(01)00018-x

Loewenberg Ball, D., & Forzani, F. M. (2009a). The Work of Teaching and the Challenge for Teacher Education. *Journal of Teacher Education*, *60*(5), 497–511. https://doi.org/10.1177/0022487109348479

Lortie, D. (1975). A sociological study. *Journal of Teacher Education*, *26*, 360-363.

MacQueen, K. (2006). Research in Review: Teacher Burnout. *Research Review Magazine, Florida State University. Retrieved from http://www. rinr. fsu. edu/issues/2006winter/_doc/fullissue. pdf*.

Marteau, T. M., & Bekker, H. (1992). The development of a six-item short-form of the state scale of the Spielberger State—Trait Anxiety Inventory (STAI). *British journal of clinical Psychology*, *31*(3), 301-306.

Mashburn, A. J., Hamre, B. K., Downer, J. T., & Pianta, R. C. (2006). Teacher and Classroom Characteristics Associated With Teachers' Ratings of Prekindergartners' Relationships and Behaviors. *Journal of Psychoeducational Assessment*, *24*(4), 367–380. https://doi.org/10.1177/0734282906290594

McCormack, A. N. N., & Thomas, K. (2003).. Is Survival Enough? Induction experiences of beginning teachers within a New South Wales context. *Asia-Pacific Journal of Teacher Education*, *31*(2), 125–138. https://doi.org/10.1080/13598660301610

McCrae, R. R., & Costa Jr, P. T. (2004). A contemplated revision of the NEO Five-Factor Inventory. *Personality and individual differences*, *36*(3), 587-596.

Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology*, *81*(2), 247–258. https://doi.org/10.1037/0022-0663.81.2.247

Mitchel, A. L., & King, M. S. (2016). A New Agenda: Research to Build a Better Teacher Preparation Program. *Bellwether Education Partners*.

Mitchell, A., & Arnold, M. (2004). Behavior Management Skills as Predictors of Retention Among South Texas Special Educators. *Journal of instructional Psychology*, *31*(3).

Munroe, A., & Pearson, C. (2006). The Munroe multicultural attitude scale questionnaire: A new instrument for multicultural studies. *Educational and Psychological Measurement*, *66*(5), 819-834.

Narvaez, D., Vaydich, J. L., Turner, J. C., & Khmelkov, V. (2008). Teacher Self-Efficacy For Moral Education. *Journal of Research in Character Education*, *6*(2).

Navidnia, H. (2009). Psychological characteristics of English language teachers: On the relationship among big five personality traits and teacher efficacy beliefs.
Neufeld, B., & Roper, D. (2003). Coaching: A strategy for developing instructional capacity.
Newman-Carlson, D., & Horne, A. M. (2004). Bully Busters: A Psychoeducational Intervention for Reducing Bullying Behavior in Middle School Students. *Journal of Counseling & Development*, *82*(3), 259–267. https://doi.org/10.1002/j.1556-6678.2004.tb00309.x
Ní Chróinín, D., & O'Sullivan, M. (2014). From initial teacher education through induction and beyond: a longitudinal study of primary teacher beliefs. *Irish Educational Studies*, *33*(4), 451–466. https://doi.org/10.1080/03323315.2014.984387
Nosek, B. A., & Errington, T. M. (2020). What is replication?. *PLoS biology*, *18*(3), e3000691.
Oliver, R. M., & Reschly, D. J. (2007). Effective Classroom Management: Teacher Preparation and Professional Development. TQ Connection Issue Paper. *National comprehensive center for teacher quality*.
O'Neill, S., & Stephenson, J. (2012). Does classroom management coursework influence pre-service teachers' perceived preparedness or confidence? *Teaching and Teacher Education*, *28*(8), 1131–1143. https://doi.org/10.1016/j.tate.2012.06.008
Ozder, H. (2011). Self-Efficacy Beliefs of Novice Teachers and Their Performance in the Classroom. *Australian Journal of Teacher Education*, *36*(5), 1–15. https://doi.org/10.14221/ajte.2011v36n5.1
Pajares, M. F. (1992). Teachers' Beliefs and Educational Research: Cleaning Up a Messy Construct. *Review of Educational Research*, *62*(3), 307–332. https://doi.org/10.3102/00346543062003307
Palomera, R., Fernández-Berrocal, P., & Brackett, M. A. (2008). Emotional intelligence as a basic competency in pre-service teacher training: Some evidence.
Papay, T. P., & Laski, M. E. (2018). Exploring teacher improvement in Tennessee: A brief on reimagining state support for professional learning. *Nashville, TN: Tennessee Education Research Alliance. Retrieved October*, *30*, 2018.
Pendergast, D., Garvis, S., & Keogh, J. (2011). Pre-Service Student-Teacher Self-efficacy Beliefs: An Insight Into the Making of Teachers. *Australian Journal of Teacher Education*, *36*(12), 46–58. https://doi.org/10.14221/ajte.2011v36n12.6
Perry, H. M. (1939). *The relative efficiency of actual and" imaginary" practice in five selected tasks* (No. 243). RS Woodworth.
Pfitzner-Eden, F. (2016). Why Do I Feel More Confident? Bandura's Sources Predict Preservice Teachers' Latent Changes in Teacher Self-Efficacy. Frontiers in Psychology, 7, 1486. https://doi.org/10.3389/fpsyg.2016.01486
Pianta, R. C. (2005). Editorial| Journal of School Psychology-Volume 43, Issue 3.

Ponnock, A. R., Torsney, B. M., & Lombardi, D. (2018). Motivational Differences throughout Teachers' Preparation and Career. *New Waves-Educational Research and Development Journal*, *21*(2), 26-45.

Poulou, M. S., Reddy, L. A., & Dudek, C. M. (2018). Relation of teacher self-efficacy and classroom practices: A preliminary investigation. *School Psychology International*, *40*(1), 25–48. https://doi.org/10.1177/0143034318798045

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R., & Du Toit, M. (2000). HLM 6: Hierarchical linear and nonlinear modeling. Scientific Software International. *Inc., Lincolnwood, IL*.

Redman, S. F. (2015). Self-efficacy and teacher retention: Perception of novice teachers on job preparation, job support, and job satisfaction.

Reisman, A., Cipparone, P., Jay, L., Monte-Sano, C., Schneider Kavanagh, S., McGrew, S., & Fogo, B. (2019). Evidence of emergent practice: Teacher candidates facilitating historical discussions in their field placements. *Teaching and Teacher Education*, *80*, 145–156. https://doi.org/10.1016/j.tate.2018.12.014

Responsive Classroom. (2014). The responsive classroom approach: Good teaching changes the future. https://www.responsiveclassroom.org/sites/default/files/pdf_files/RC_approach White_paper.pdf

Richards, J. C., & Lockhart, C. (1996). Exploring teachers' beliefs. *JC Richards, & C. Lockhart, Reflective Teaching in Second Language Classrooms*, 29-51.

Riddle, R. L. (2018). Examining the Relationship between Pre-Service Teachers' Self-Efficacy Beliefs and Grit Levels.

Rimm-Kaufman, S. E., & Hamre, B. K. (2010). The role of psychological and developmental science in efforts to improve teacher quality. *Teachers College Record*, *112*(12), 2988-3023.

Ripski, M. B., LoCasale-Crouch, J., & Decker, L. (2011). Pre-service teachers: Dispositional traits, emotional states, and quality of teacher-student interactions. *Teacher Education Quarterly*, *38*(2), 77-96.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, Schools, and Academic Achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rock, M. L., Schoenfeld, N., Zigmond, N., Gable, R. A., Gregg, M., Ploessl, D. M., & Salter, A. (2013). Can you Skype me Now? Developing Teachers' Classroom Management Practices through Virtual Coaching. *Beyond Behavior*, *22*(3), 15–23. https://doi.org/10.1177/107429561302200303

Rockoff, J. E. (2004). The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data. *American Economic Review*, *94*(2), 247–252. https://doi.org/10.1257/0002828041302244

Roness, D. (2011). Still motivated? The motivation for teaching during the second year in the profession. *Teaching and Teacher Education*, *27*(3), 628–638. https://doi.org/10.1016/j.tate.2010.10.016

Ronfeldt, M. (2015). Field Placement Schools and Instructional Effectiveness. *Journal of Teacher Education*, *66*(4), 304–320. https://doi.org/10.1177/0022487115592463

Rosas, C., & West, M. (2009). Teachers beliefs about classroom management: Pre-service and inservice teachers' beliefs about classroom management. International Journal of Applied Educational Studies, *5(*1), 54-62.

Ross, S. L. (1985). The Effectiveness of Mental Practice in Improving the Performance of College Trombonists. *Journal of Research in Music Education*, *33*(4), 221–230. https://doi.org/10.2307/3345249

Rubin, J., & Herbert, M. (1998). Peer teaching-model for active learning. *College Teaching*, *48*(1), 26-30.

Sabar, N. (2004). From heaven to reality through crisis: novice teachers as migrants. *Teaching and Teacher Education*, *20*(2), 145–161. https://doi.org/10.1016/j.tate.2003.09.007

Sanders, C. W., Sadoski, M., van Walsum, K., Bramson, R., Wiprud, R., & Fossum, T. W. (2008). Learning basic surgical skills with mental imagery: using the simulation centre in the mind. *Medical Education*, *42*(6), 607–612. https://doi.org/10.1111/j.1365-2923.2007.02964.x

Savran, A., & Çakiroglu, J. (2003). Differences between Elementary and Secondary Pre-service Science Teachers' Perceived Efficacy Beliefs and Their Classroom Management Beliefs. *Turkish Online Journal of Educational Technology-TOJET*, *2*(4), 15-20.

Scheeler, M. C., McAfee, J. K., Ruhl, K. L., & Lee, D. L. (2006). Effects of Corrective Feedback Delivered via Wireless Technology on Preservice Teacher Performance and Student Behavior. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, *29*(1), 12–25. https://doi.org/10.1177/088840640602900103

Scheeler, M. C. (2007). Generalizing Effective Teaching Skills: The Missing Link in Teacher Preparation. *Journal of Behavioral Education*, *17*(2), 145–159. https://doi.org/10.1007/s10864-007-9051-0

Scheeler, M. C., Bruno, K., Grubb, E., & Seavey, T. L. (2009). Generalizing teaching techniques from university to K-12 classrooms: Teaching pre-service teachers to use what they learn. *Journal of Behavioral Education*, *18*(3), 189-210.

Schiefele, U., & Schaffner, E. (2015). Teacher interests, mastery goals, and self-efficacy as predictors of instructional practices and student motivation. *Contemporary Educational Psychology*, *42*, 159–171. https://doi.org/10.1016/j.cedpsych.2015.06.005

Seiz, J., Voss, T., & Kunter, M. (2015). When Knowing Is Not Enough--The Relevance of Teachers' Cognitive and Emotional Resources for Classroom Management. *Frontline Learning Research*, *3*(1), 55-77.

Shaefer, E. S., & Edgerton, M. (1985). Parent and child correlates of parental modernity. W: IE Sigel (red.), Parental belief systems (s. 287–318).

Shoffner, M. (2008). Informal reflection in pre-service teacher education. *Reflective Practice*, *9*(2), 123–134. https://doi.org/10.1080/14623940802005392

Skaalvik, E. M., & Skaalvik, S. (2017). Motivated for teaching? Associations with school goal structure, teacher self-efficacy, job satisfaction and emotional exhaustion. *Teaching and Teacher Education*, *67*, 152-160.

Skinner, E., & Beers, J. (2016). Mindfulness and teachers' coping in the classroom: A developmental model of teacher stress, coping, and everyday resilience. In *Handbook of mindfulness in education* (pp. 99-118). Springer, New York, NY.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549–3557. https://doi.org/10.1098/rstb.2009.0138

Sockett, H. (2009). Dispositions as Virtues. *Journal of Teacher Education*, *60*(3), 291–303. https://doi.org/10.1177/0022487109335189

Soodak, L. C., Podell, D. M., & Lehman, L. R. (1998). Teacher, Student, and School Attributes as Predictors of Teachers' Responses to Inclusion. *The Journal of Special Education*, *31*(4), 480–497. https://doi.org/10.1177/002246699803100405

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Stai. *Manual for the State-Trait Anxiety Inventory (Self Evaluation Questionnaire). Palo Alto California: Consulting Psychologist*, *22*, 1-24.

Stahl, G., Sharplin, E., & Kehrwald, B. (2016). Developing pre-service teachers' confidence: real-time coaching in teacher education. *Reflective Practice*, *17*(6), 724–738. https://doi.org/10.1080/14623943.2016.1206882

Stapleton, J., Tschida, C., & Cuthrell, K. (2017). Partnering principal and teacher candidates: Exploring a virtual coaching model in teacher education. *Journal of Technology and Teacher Education*, *25*(4), 495-519.

Steele, J. L., Pepper, M. J., Springer, M. G., & Lockwood, J. (2015). The distribution and mobility of effective teachers: Evidence from a large, urban school district. *Economics of Education Review*, *48*, 86–101. https://doi.org/10.1016/j.econedurev.2015.05.009

Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance. *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

Steiner, L., & Kowal, J. (2007). Instructional coaching. *Reading Rockets*.

Steiner, P. M., & Wong, V. C. (2018). Assessing Correspondence Between Experimental and Nonexperimental Estimates in Within-Study Comparisons. *Evaluation Review*, *42*(2), 214–247. https://doi.org/10.1177/0193841x18773807

Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2010). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 369–386. https://doi.org/10.1111/j.1467-985x.2010.00673.x

Sutton, R. E. (2007). Teachers' anger, frustration, and self-regulation. In *Emotion in education* (pp. 259-274). Academic Press.

Tipton, E., & Olsen, R. B. (2018). A Review of Statistical Methods for Generalizing From Evaluations of Educational Interventions. *Educational Researcher*, *47*(8), 516–524. https://doi.org/10.3102/0013189x18781522

Topoğlu, O. (2014). Musical Performance Anxiety: Relations between Personal Features and State Anxiety Levels of Pre-Service Music Teachers. *International Online Journal of Educational Sciences*, *6*(2), 337–348. https://doi.org/10.15345/iojes.2014.02.008

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher Efficacy: Its Meaning and Measure. *Review of Educational Research*, *68*(2), 202–248. https://doi.org/10.3102/00346543068002202

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783–805. https://doi.org/10.1016/s0742-051x(01)00036-1

Tschannen-Moran, M., & Hoy, A. W. (2007). The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education*, *23*(6), 944–956. https://doi.org/10.1016/j.tate.2006.05.003

Veenman, S. (1984). Perceived Problems of Beginning Teachers. *Review of Educational Research*, *54*(2), 143–178. https://doi.org/10.3102/00346543054002143

Villaume, S. K. (2000). The Necessity of Uncertainty: A Case Study of Language Arts Reform. *Journal of Teacher Education*, *51*(1), 18–25. https://doi.org/10.1177/002248710005100103

Vogt, F., & Rogalla, M. (2009). Developing Adaptive Teaching Competency through coaching. *Teaching and Teacher Education*, *25*(8), 1051–1060. https://doi.org/10.1016/j.tate.2009.04.002

Voss, T., & Kunter, M. (2019). "Reality Shock" of Beginning Teachers? Changes in Teacher Candidates' Emotional Exhaustion and Constructivist-Oriented Beliefs. *Journal of Teacher Education*, *71*(3), 292–306. https://doi.org/10.1177/0022487119839700

Wagler, R. (2011). The impact of vicarious experiences and field experience classroom characteristics on pre-service elementary science teaching efficacy. *Electronic Journal of Science Education*, *15*(2), 1-28.

Waschbusch, D. A., & Willoughby, M. T. (2007). Parent and Teacher Ratings on the IOWA Conners Rating Scale. *Journal of Psychopathology and Behavioral Assessment*, *30*(3), 180–192. https://doi.org/10.1007/s10862-007-9064-y

Wenner, G. (2001). Science and mathematics efficacy beliefs held by practicing and prospective teachers: A 5-year perspective. Journal of Science Education and Technology, 10(2), 181–187. https://doi.org/10.1023/A:1009425331964

Wetzel, C. M., George, A., Hanna, G. B., Athanasiou, T., Black, S. A., Kneebone, R. L., Nestel, D., & Woloshynowych, M. (2011). Stress Management Training for Surgeons—A Randomized, Controlled, Intervention Study. *Annals of Surgery*, *253*(3), 488–494. https://doi.org/10.1097/sla.0b013e318209a594

Wiens, P. D., & Ruday, S. (2014). Personality and Pre-service Teachers: Does It Change, Does It Matter?. *Issues in Teacher Education*, *22*(2), 7-27.

Wilson, S. M., Shulman, L. S., & Richert, A. E. (1987). One hundred fifty different ways of knowing: Representations of knowledge in teaching. *Exploring teachers' thinking*, 104-124.

Wilson, S. M., Floden, R., & Ferrini-Mundy, J. (2001). Teacher preparation research: Current knowledge, gaps. *CTP Center for the Study of Teaching and Policy Center for the Study of Teaching and Policy Teachers College/Columbia University*.

Windschitl, M. (2002). Inquiry projects in science teacher education: What can investigative experiences reveal about teacher thinking and eventual classroom practice? *Science Education*, *87*(1), 112–143. https://doi.org/10.1002/sce.10044

Wong, V. C., & Steiner, P. M. (2018). *Replication designs for causal inference*. EdPolicyWorks Working Paper Series. Retrieved from http://curry. Virginia. edu/uploads/epw/62_Replication_Designs. pdf.

Wong, V. C., Steiner, P. M., & Anglin, K. L. (2020). Design-BasedApproachestoSystematicConceptual ReplicationStudies. EdPolicyWorks Working Paper Series.

Woolfolk, A. E., Rosoff, B., & Hoy, W. K. (1990). Teachers' sense of efficacy and their beliefs about managing students. *Teaching and Teacher Education*, *6*(2), 137–148. https://doi.org/10.1016/0742-051x(90)90031-y

Wyatt, M. (2015). "Are they becoming more reflective and/or efficacious?" A conceptual model mapping how teachers' self-efficacy beliefs might grow. *Educational Review*, *68*(1), 114–137. https://doi.org/10.1080/00131911.2015.105875

Yokuş, T. (2012). The relationship between the state-trait anxiety levels and the academic achievement of music teacher candidates. *International Online Journal of Primary Education (IOJPE) ISSN: 1300-915X*, *2*(1).

Yost, D. S. (2006). Reflection and self-efficacy: enhancing the retention of qualified teachers from a teacher education perspective. Teacher Education Quarterly, 33, 59 –76 Retrieved from https://www.jstor.org/stable/23478871

Zee, M., & Koomen, H. M. Y. (2016). Teacher Self-Efficacy and Its Effects on Classroom Processes, Student Academic Adjustment, and Teacher Well-Being. *Review of Educational Research*, *86*(4), 981–1015. https://doi.org/10.3102/0034654315626801

Zeichner, K. (2014). The struggle for the soul of teaching and teacher education in the USA. *Journal of Education for Teaching*, 1–18. https://doi.org/10.1080/02607476.2014.956544

Zembylas, M. (2007). Emotional ecology: The intersection of emotional knowledge and pedagogical content knowledge in teaching. *Teaching and Teacher Education*, *23*(4), 355–367. https://doi.org/10.1016/j.tate.2006.12.002

Appendix Table 1

Causal Replication Framework for the Direct Replication of Effects

Design Assumptions	For Study 1 …	… Through Study *k*
Replication assumptions (*R1*-*R2*)	R1. Treatment and outcome stability R2. Equivalence in the causal estimand	
Individual study assumptions (*S1*-*S3*)	S1. Unbiased identification of effects S2. Unbiased estimation of effects S3. Correct reporting of estimators, estimands, and estimates	S1. Unbiased identification of effects S2. Unbiased estimation of effects S3. Correct reporting of estimators, estimands, and estimates

Source. Steiner, Wong, & Anglin, 2020; Wong & Steiner, 2018

Appendix Table 2

CRF Assumptions Tested in Planned Causal Replication Study

	R1. Treatment/ Outcome Stability	R2. Equivalent Causal Estimand	S1. Identification	S2. Estimation	S3. Reporting
Switching Replication Design (Study 2 vs. Study 3)	Treatments ✓ Outcomes ✓	Participants ✓ Settings × Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Multiple Cohort Design (Study 1 vs. Study 3)	Treatments ✓ Outcomes ✓	Participants ✓ Settings ✓ Causal quantity ✓ Time ×	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Conceptual Replication Design with different training contexts (Study 3 vs. Study 4)	Treatments ✓ Outcomes ✓	Participants ✓ Settings × Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Conceptual Replication Design with different modes of delivery (Study 3 vs. Study 5)	Treatments ✓ Outcomes ✓	Participants ✓ Settings × Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓

Appendix A. Propensity Score Weighting for Replication Study 3 and Study 4

Procedures used for Estimating Propensity Weights

Across studies 3 and 4, there are four groups of participants. We note each of the four groups using two indices: the first is the study index (with values 0 for the group of teacher candidates in study 3 and 1 for the group of undergraduate students in study 4) and the second index is assignment to the coaching condition (noted as 1) or self-reflection condition (noted as 0). The four resulting groups as follows:

Y (0,0), teacher candidates in study 3 randomly assigned to self-reflection

Y (0,1), teacher candidates in study 3 randomly assigned to coaching

Y (1,0), undergraduate participants in study 3 randomly assigned to self-reflection

Y (1,1), undergraduate participants in study 3 randomly assigned to coaching

Given a set of N observations where each observation is coded as a member of one of the four groups [(0,0); (0,1); (1,0) (1,1)], we fit a multinomial regression model where the probabilities of membership in each of the four groups is as follows:

$$\Pr\left(M_i = (0,0)\right) = 1 - (\Pr\left(M_1 = (0,1)\right) + \Pr\left(M_i = (1,0)\right) + \Pr\left(M_1(1,1)\right)$$

$$\Pr\left(M_i = (0,1)\right) = \frac{\exp\left[X_i\beta_{(0,1)}\right]}{1 + \exp\left[X_i\beta_{(0,1)}\right] + \exp\left[X_i\beta_{(1,0)}\right] + \exp\left[X_i\beta_{(1,1)}\right]}$$

$$\Pr\left(M_i = (1,0)\right) = \frac{\exp\left[X_i\beta_{(1,0)}\right]}{1 + \exp\left[X_i\beta_{(0,1)}\right] + \exp\left[X_i\beta_{(1,0)}\right] + \exp\left[X_i\beta_{(1,1)}\right]}$$

$$\Pr\left(M_i = (1,1)\right) = \frac{\exp\left[X_i\beta_{(1,1)}\right]}{1 + \exp\left[X_i\beta_{(0,1)}\right] + \exp\left[X_i\beta_{(1,0)}\right] + \exp\left[X_i\beta_{(1,1)}\right]}$$

Where each is a set of coefficients that relate the matrix of covariates, X , to group membership.

The weight for each case was then defined as the inverse of the estimated probability of membership to the case's observed group as given below:

$$w_i \begin{cases} \frac{1}{\Pr(M_i = (0,0))} \; if \; M_i = (0,0) \\ \frac{1}{\Pr(M_i = (0,1))} \; if \; M_i = (0,1) \\ \frac{1}{\Pr(M_i = (1,0))} \; if \; M_i = (1,0) \\ \frac{1}{\Pr(M_i = (1,1))} \; if \; M_i = (1,1) \end{cases}$$

Next, we estimate the means for each group with a weighted regression that also includes the covariates X, Y(M) using Stata's teffects command with the inverse propensity weighting and regression adjustment (ipwra) procedure testing weighted potential outcome means. For the propensity weighting process, we used several covariates, including indicator variables for whether participants were female, over the age of 21, White, attended an urban high school, attended a high school where students were high achieving, attended a high school average SES and continuous variables for participants' baseline self-efficacy beliefs, multicultural attitudes towards teaching and beliefs regarding culturally-responsive teaching. These covariates were also included in the regression model used to produce weighted means.

Results of Propensity Weighting

Table B1 displays standardized differences raw between the undergraduate sample in Study 3 and teacher candidate sample in Study 4 prior to and after propensity score weighting, using the teacher candidate sample as the reference group. Across the covariates, standardized differences in just four (% age above 21, % of participants who attended an urban high school), overall self-efficacy beliefs and average culturally responsive teaching self-efficacy beliefs fall within the acceptable magnitude of 0.10. All raw standardized differences for the remaining covariates are 0.10 or greater. Using propensity score weighting, all standardized differences of covariates in Table B1 are less than 0.03 and there acceptable by convention.

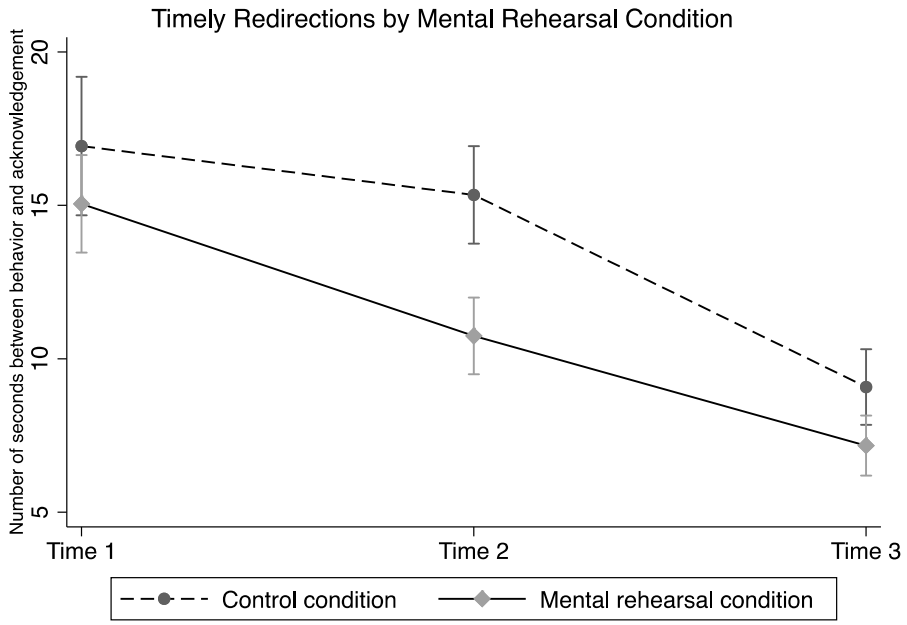
Appendix Table 3

Standardized Raw Differences between Study 3 and Study 4

	Before weighting	After weighting
	Standardized raw difference	
% Female	0.16	-0.02
% Over the age of 21	0.05	0.02
% White	-0.12	-0.02
Location of high school attended		
% Urban	-0.08	-0.02
Average SES of high school attended		
% Middle SES	-0.17	0.00
Average achievement level of high school attended		
% Primarily high achieving	-0.18	0.00
Self-efficacy Beliefs (overall)	0.09	0.03
Self-efficacy Beliefs (classroom management)	0.10	0.03
Teacher Multicultural Attitudes score	-0.13	-0.02
Culturally Responsive Teaching Self-efficacy score	0.09	0.01

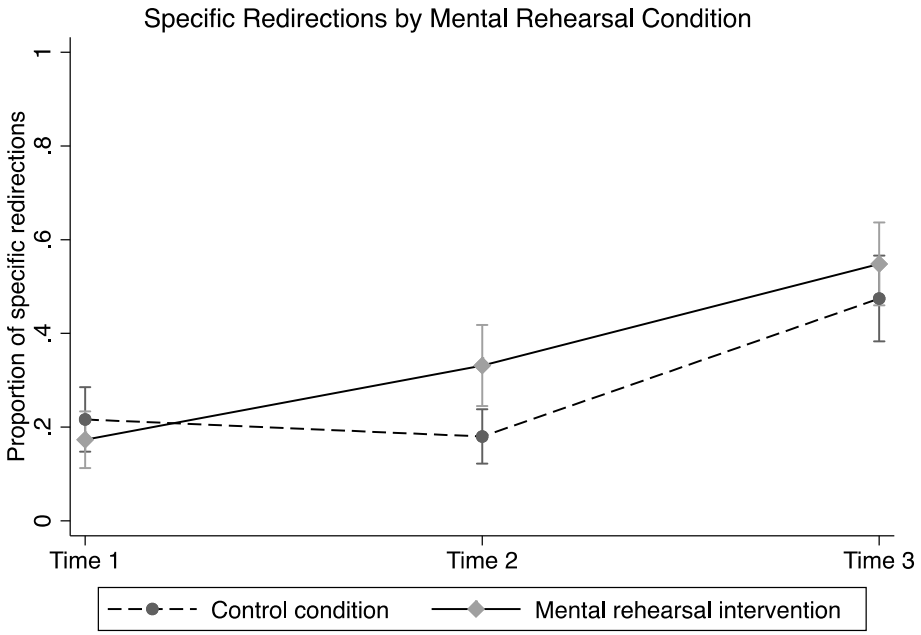
Appendix Figure 1.

Improvements in Timely Redirections by Mental Rehearsal Condition



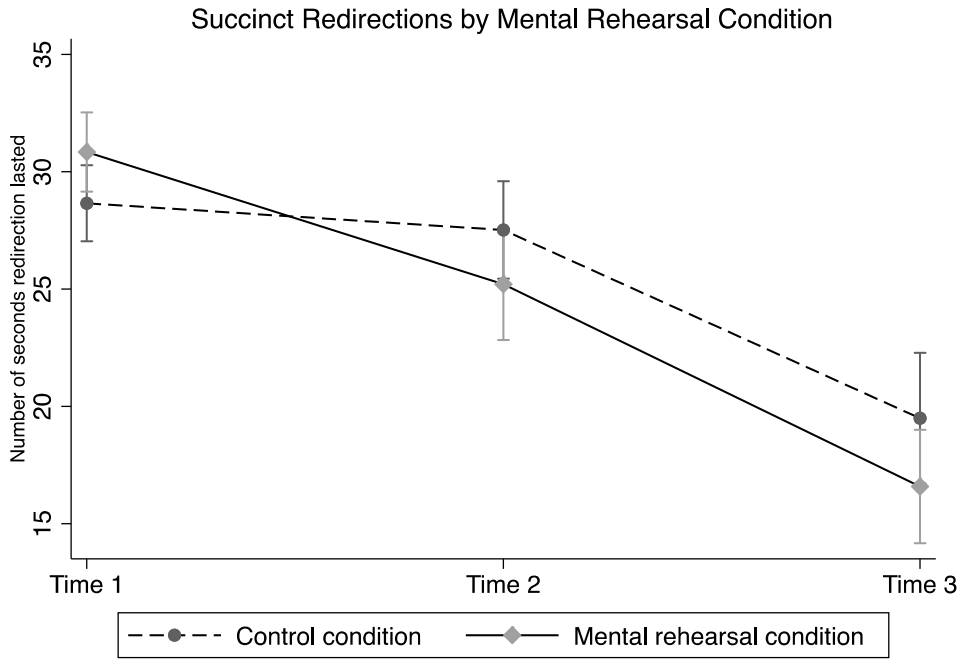
Appendix Figure 2.

Improvements in Proportion of Specific Redirections by Mental Rehearsal Condition



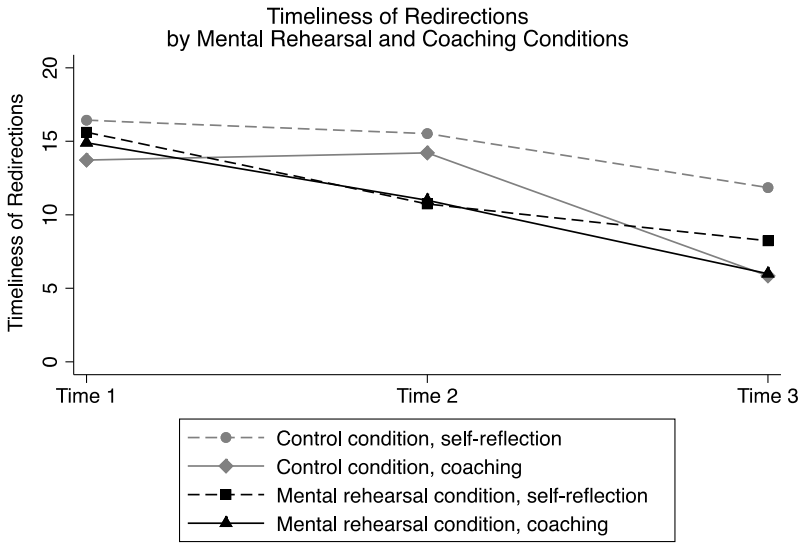
Appendix Figure 3.

Improvements in Succinctness of Redirections by Mental Rehearsal Condition



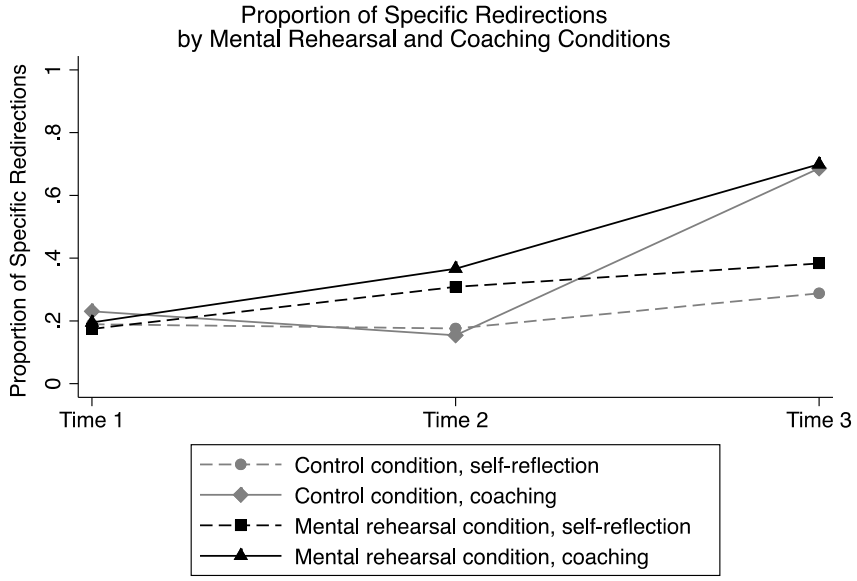
Appendix Figure 4.

Improvements in Timeliness of Redirections by Mental Rehearsal and Coaching Conditions



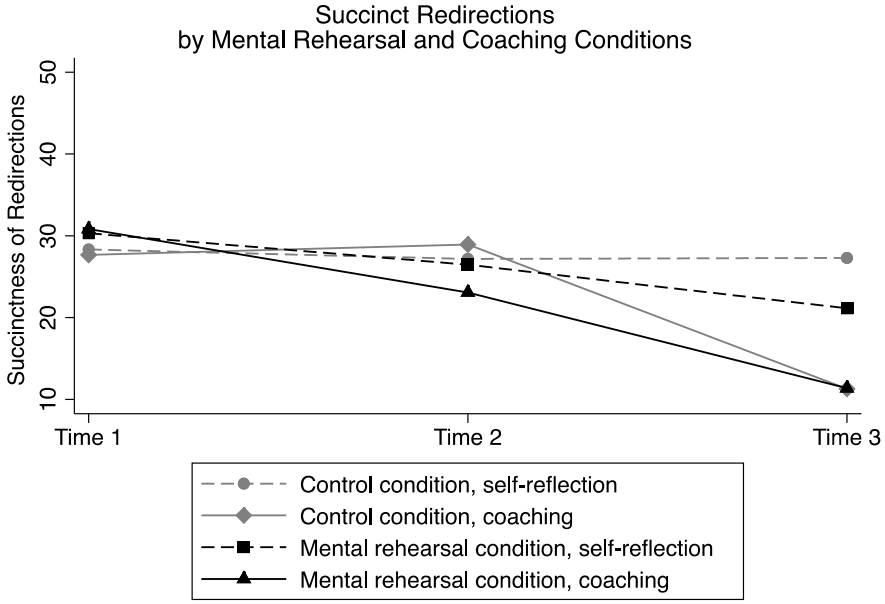
Appendix Figure 5.

Improvements in Proportion of Specific Redirections by Mental Rehearsal and Coaching Conditions



Appendix Figure 6.

Improvements in Succinctness of Redirections by Mental Rehearsal and Coaching Conditions



Appendix Table 4.

Cross-sectional results of impact of mental rehearsal on teacher candidates' outcomes at time 2

	Regression-Adjusted Control Group	Mental Rehearsal/ Control difference						
	Mean (SD) (1)	**Coefficient (SE) (2)**	**Coefficient (SE) (3)**	**Coefficient (SE) (4)**	**Coefficient (SE) (5)**	**Coefficient (SE) (6)**	**Glass' Delta**	**N**
Coded Performance Outcomes								
Overall Quality Score	3.95	0.79**	0.79**	0.78**	0.80**	0.73**	0.65	105
	(0.51)	(0.25)	(0.25)	(0.26)	(0.26)	(0.25)		
Timeliness of Redirections	13.90	-3.87**	-3.87**	-3.93**	-3.68**	-3.85**	-0.77	105
	(2.20)	(0.92)	(0.92)	(0.97)	(0.97)	(0.93)		
Proportion of Specific Redirections	0.18	0.14*	0.14*	0.13*	0.13*	0.15**	0.65	105
	(0.12)	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)		
Succinctness of Redirections	28.99	-3.22*	-3.21*	-2.60	-3.28*	-2.81+	-0.47	105
	(3.73)	(1.60)	(1.60)	(1.64)	(1.61)	(1.54)		
Pre-simulation Survey Outcomes								
Self-efficacy in Classroom Management	6.24	-0.39*	-0.40*	-0.38+	-0.26	-0.28	-0.44	105
	(0.39)	(0.20)	(0.20)	(0.20)	(0.18)	(0.18)		
STAI- Totaled	15.01	0.29	0.28	0.56	-	0.53	0.06	105
	(1.75)	(0.69)	(0.68)	(0.69)	-	(0.69)		
Randomization blocks		x	x	x	x	x		
Coaching condition			x	x	x	x		
Covariates				x	x	x		
Baseline outcomes					x	x		
Interactor controls						x		

Coach controls						x

Note. Adjusted control group means and clustered standard errors (in parentheses) are reported in Column (1). Coefficients and clustered standard errors are reported in Columns (2) through (6) represent mean adjusted differences between control and mental rehearsal conditions. Coefficients are taken from separate regressions of the outcome on intervention assignment. Model 1 includes controls for randomization blocks, Model 2 adds controls for whether teacher candidate was above the age of 21 at baseline and pretest outcome measures including an overall quality score, proportion of specific redirections, timeliness of redirections and succinctness of redirections. I include missing data indicators in cases where respondents have missing pre-treatment covariates. Glass' Delta was calculated using the mean difference between the control group and mental rehearsal group, and then dividing the result by the standard deviation of the control group †p < .10. *p < .05. **p < .01.

Appendix B: Mental Rehearsal Script

Below is a quick overview of the classroom norms and expectations task to jog your memory on what to expect in the scenario. As you read the overview, it would be helpful to imagine that you are in the simulator and use your memory of the summer/fall session you completed to help you follow the script.

Imagine that you are practicing the classroom norms scenario. At the start of the simulator session, take a couple of seconds for introductions. Since it's the first day of school, this will give you and the student avatars a chance to know each other before diving into the discussion. Keep the introduction brief to spend as much time on the norm setting discussion as possible. You may also choose to skip introductions, if you already feel comfortable beginning the discussion with the student avatars.

Explain to the student avatars that today you will be having a discussion on setting class norms and expectations that help you work as productive learning community. You should also make clear that the student avatars will decide on these norms collaboratively.

As you facilitate the discussion, remember to pay attention to how student avatars in the classroom are behaving, in addition to moving the norm-setting discussion forward. If a student is off-task, bring them back to the discussion by responding to

them in a timely manner. The sooner you address off-task behavior, the sooner you will be able to re-engage the student avatar in the classroom discussion and prevent the behavior from escalating.

The best way to stop off-task behavior in the classroom is to provide the student with clear, concise, and specific instructions about what they are doing that is off-task and how they can engage effectively in the activity. You can explicitly tell the student avatar to stop the particular off-task behavior, as well as tell them what kind of behavior you would like them to display. Use as few words as possible so you can avoid detracting from the norm-setting discussion and drawing attention to the off-task student behavior. The goal is to help the student understand exactly what you are asking them to do and to bring them back to the discussion immediately.

Although it can be frustrating when students engage in off-task behaviors, remember to remain warm but firm while redirecting the off-task behavior. Doing so will ensure a positive classroom environment, prevent the off-task behavior from escalating and not take time away from the collaborative discussion.

Your goal is for the class to generate at least three behavioral expectations and provide rationale for each by the end of the five-minute simulation. Remember that while it's important to keep the discussion brief and focused so that you can generate the three

expectations, you also want to make sure the students get a chance to talk about why these specific expectations will help make the classroom productive and safe.

You will have five minutes to practice in the simulator, following which there will be a five-minute break. During this break, you will either complete a self-reflection protocol or receive feedback from a coach. As a reminder, regardless of whether or not you complete a self-reflection protocol or receive feedback from a coach, we are not evaluating your performance! We are just trying to understand the best ways to support teacher candidates' learning in a simulated setting.

Don’t be nervous, the next simulation will feel a lot easier since you have some practice from the summer/fall. Good luck!

Appendix C: Structural Equation Models

In addition to fitting regression models to estimate the causal impact of mental rehearsal on teacher candidates' performance and perception outcomes in the simulator, I also use a structural equation modeling framework. SEM models allow for repeated measures of my variables of interest, to study direct and indirect effects of mental rehearsal on teacher candidate outcomes, and most importantly, to account for measurement error using latent variables (Bollen & Noble, 2011; Kline, 2011).

Model specification and refinements. The initial structure model hypothesized in Appendix Figure 7 was analyzed using SEM in STATA. The initial SEM model was based on theoretical expectations the regression models presented in Equations 1, 2 and 3. Using Goodness-Of-Fit (GOF) measures, model refinement was performed to improve the fit to recommended levels. The final model fitting for teacher candidates' stress levels, self-efficacy beliefs and performance outcomes based on essential GOF measures is adequately supported. Although the model has a significant chi-squared, both the root mean squared error of approximation (RMSEA) and comparative fit index (CFI) are excellent. Specifically, the RMSEA value of 0.051 suggests that the final model cannot be rejected at the 95% confidence interval. The CFI value of 0.95 also provides strong evidence that the fit between the measurement model and the data is acceptable (Jin et al., 2007; Molenaar and Washington, 2000; Wood and Ellis, 2005).

Reliability of constructs. As highlighted in the text in the main paper and presented in Table 12, I used Cronbach's alpha to determine the reliability of each of the

constructs used in the SEM model. For Cronbach's alpha, a cut-off value of 0.7 indicates acceptable level of internal consistency. As seen from Table 12, the attributes measuring stress, self-efficacy beliefs and behavior ratings in the final SEM model resulted in high degree of reliability above the cut-off value of 0.7.

Results of SEM model. Figure B1 depicts the final cross-sectional model used to examine the impact of mental rehearsal on participants' performance and perception outcomes after deleting paths that are not of interest in this paper. As seen, all of the path coefficients from the SEM model trend in the same direction as the regression model results, and they follow the same statistical significance patterns.

Teacher candidates' performance and perceptions at time 2. The SEM results (presented in Figure B1) suggest that participants who completed the mental rehearsal protocol prior to simulation sessions at Time 2 saw an increase in their stress levels (approximately 0.6 points on the STAI measure) compared to participants in the control condition, as well as a slight decrease in their self-efficacy beliefs in classroom management (about 0.24 points). In addition, candidates in the mental rehearsal condition saw a significant improvement in their instructional skills in the simulator at Time 2, immediately after they completed the mental rehearsal protocol. Specifically, mental rehearsal significantly increased participants' overall quality score by approximately 0.71 points, improved the timeliness of redirections so that candidates reacted about 4 seconds faster to student avatar behaviors than their peers in the control group and increased the proportion of specific redirections that teacher candidates provided in reaction to student

avatar behaviors (an increase of about 17%). In addition, teacher candidates in the mental rehearsal condition provided more succinct redirections in that their redirections lasted approximately 3 seconds lesser than redirections provided by control group participants, although this difference is not statistically significant.

Teacher candidates' performance and perceptions at time 3. At Time 3 (after participants have either completed the self-reflection protocol or received coaching depending on the condition to which they were randomly assigned), Figure B1 shows that mental rehearsal did not have a significant impact on any of the outcomes of interest (either performance or perception). Although mental rehearsal did continue to influence participants' performance outcomes in that they scored higher on the measure of overall quality score by 0.07 points, provided 13% more specific redirection and took approximately 2 seconds lesser on average to redirect student avatar behavior, none of these coefficients were significant with a p-value <0.10. From Figure B1, it also appears that while mental rehearsal did not have a significant impact on either participants' self-efficacy beliefs in classroom management or their endorsement of punitive management approaches at Time 2, participants in the mental rehearsal condition did rate student avatars as displaying oppositional/defiant behaviors to a significantly lesser extent (about 0.23 points).

Overall the results support the hypothesis that mental rehearsal is a useful tool at improving participants' instructional performance in the simulator, and corroborate results from the regression models presented in Equations 1-3. At Time 2, immediately

after participants who were randomly assigned to the mental rehearsal condition completed the protocol, there was a significant and positive relationship between mental rehearsal and performance outcomes including overall quality score (coefficient=0.71, $p<0.01$), timeliness of redirections (coefficient=-3.98, $p<0.01$), proportion of specific redirections (coefficient=0.17, $p<0.01$) and succinctness of redirections (coefficient=-2.45, $p>0.10$). This finding supports hypothesis 2. However, the lack of association between mental rehearsal and self-efficacy beliefs (coefficient=-0.24, $p>0.10$) and mental rehearsal stress (coefficient=0.06, $p>0.10$) suggests that mental rehearsal did not effectively reduce stress or improve teacher candidates' confidence in their own ability to manage a disruptive classroom environment.

At Time 3, results suggest that the positive impact of mental rehearsal on teacher candidates' performance outcomes "watered down" in that although participants who completed the mental rehearsal protocol scored higher on the overall measure of quality, provided more timely and succinct redirections and provided a greater proportion of specific redirections, these associations were not statistically significant and were likely because some participants in the mental rehearsal condition also received five minutes of face-to-face coaching in between simulator sessions at Time 2 and Time 3.

Appendix Figure 7

Final Structural Equation Model