Experimental Evidence on the Robustness of Coaching Supports in Teacher Education

Julie Cohen[1], Vivian C Wong[1], Anandita Krishnamachari & Steffen Erickson

University of Virginia

January 2023

Accepted in August 2023 for publication at Educational Researcher

[1] First two authors listed alphabetically

Acknowledgements: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B140026 and Grant #R305D190043 to the Rectors and Visitors of the University of Virginia, the National Academy of Education/Spencer Foundation post-doctoral fellowship, the Jefferson Trust through Grant #DR02951 and the Bankard Fund through Grant #ER00562. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors wish to thank Peter Steiner, James Pustejovsky, Elizabeth Tipton, Jim Soland, Kylie Anglin, Emily Wiseman, Alexis Prijoles, Rose Sebastian, Christina Taylor, and the members of the TeachSIM lab at the University of Virginia for their feedback on earlier versions of this paper. All errors are those of the authors.

Abstract

Many novice teachers learn to teach "on-the-job," leading to burnout and attrition among teachers and negative outcomes for students in the long term. Pre-service teacher education is tasked with optimizing teacher readiness, but there is a lack of causal evidence regarding effective ways for preparing new teachers. In this paper, we use a mixed reality simulation platform to evaluate the causal effects and robustness of an individualized, directive coaching model for candidates enrolled in a university-based teacher education program, as well as for undergraduates considering teaching as a profession. Across five conceptual replication studies, we find that targeted, directive coaching significantly improves candidates' instructional performance during simulated classroom sessions, and that coaching effects are robust across different teaching tasks, study timing, and modes of delivery. However, coaching effects are smaller for a sub-population of participants not formally enrolled in a teacher preparation program. These participants differed from teacher candidates in multiple ways, including by demographic characteristics, as well as by their prior experiences learning about instructional methods. We highlight implications for research and practice.

Introduction

There is considerable evidence that teachers improve dramatically in their early years of classroom experience (Atteberry et al., 2015; Harris & Sass, 2011; Kraft & Papay, 2014). This on-the-job learning is stressful for beginning teachers, and the majority report entering the classroom feeling underprepared, leading to burnout, attrition, and negative outcomes for students (Ingersoll, 2001; Papay & Laski, 2018). A central question for the field has been whether—and how—we could move some of this rapid skill development into pre-service teacher education, *before* teachers become solely responsible for students. Teachers who start their careers with a solid foundation in critical instructional skills would be better poised to stay in the classroom and contribute to positive student outcomes. Unfortunately, we lack robust, causal evidence about methods for promoting this kind of rapid skill development during pre-service preparation.

Given that teachers get better "with practice," a potential avenue for development is having pre-service teachers (termed 'candidates') repeatedly practice teaching skills, with feedback and support (Grossman et al., 2009; Hoffman, et al., 2015). Traditionally, candidates are intended to practice these skills during their clinical placements, working alongside experienced mentor teachers. Unfortunately, there are clear downsides to sole reliance on this apprenticeship model. Candidates do not always have chances to practice all skills they need as teachers of record. Mentors also vary in the degree to which they model strong teaching (Ronfeldt, 2015) and may not provide necessary feedback (Matsko et al., 2020). Thus, preparation programs have been studying whether 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 candidates' ability to translate theoretical

knowledge about "effective teaching" into practice (Grossman et al., 2009b; Kavanagh & Rainey, 2017; Reisman et al., 2019). However, little work has looked at the causal effects of such approximations, or whether different supports surrounding approximations enhance their utility (Authors, 2020).

Coaching is a promising option for expediting skill development during approximations (Kraft et al., 2018). Theory suggests that coaches serve as experts who can observe teachers, evaluate strengths and weaknesses, and develop individualized strategies to promote improvement (Coburn & Woulfin, 2012). Coaching is increasingly common for in-service teachers (Cilliers et al., 2020; Majerowicz & Montero, 2018; Stahl et al., 2016) and has been shown to improve teachers' attitudes toward teaching, feelings of self-efficacy, instructional skills, and student achievement (Castro et al., 2019; Desimone & Pak, 2017; Kretlow & Bartholomew, 2010).

Despite this compelling evidence, coaching is under-utilized during pre-service preparation. Though mentors in clinical placements sometimes provide directive coaching and feedback, we often ask candidates to learn by observation and osmosis (Matsko et al., 2020). Given the short duration of teacher preparation, more standardized, frequent, and explicit feedback on developing skills could be a powerful and efficient complement to more variable clinical placements. Although the literature on pre-service coaching is nascent, a handful of studies associate coaching with improvements in candidates' satisfaction with preparation and skill development (e.g., Albornoz et al., 2020; Bowman & McCormick, 2000). We theorize that coaching could be especially useful earlier in a teacher's development when skills are only emergent, and ideas about effective practice are less ossified (Ericsson & Pool, 2016).

In this study, we examine the immediate causal effects of short, five-minute coaching sessions delivered in the context of mixed-reality simulations (MRS). The simulation platform features a virtual classroom and student avatars who are remotely controlled by an actor trained to facilitate realistic classroom interactions. The simulation sessions are integrated with candidates' coursework and serve as both a practice space and assessment platform in our research. Designed to be complementary to candidates' field placements, the simulation sessions provide candidates with opportunities to "try out" new teaching skills before practicing them in live classroom settings. In prior work, we found compelling experimental evidence that short, five-minute, directive coaching sessions can dramatically improve candidates' classroom management skills as observed in the MRS (Author, 2020). What was less clear, however, is whether such findings would replicate in additional experimental evaluations with systematic variations in participant characteristics (units), pedagogical outcomes, settings, and times (Cronbach, 1982).

To examine the efficacy and robustness of a standardized coaching protocol on candidates' instructional performance in the MRS, we conduct a series of conceptual replication studies in which we randomize candidates to receive coaching or to self-reflect on their performance in the MRS. The replication studies were designed to introduce systematic sources of variation across studies to evaluate the robustness of coaching effects across different time periods, teaching tasks and associated outcomes, participant characteristics and course experiences, and delivery modes.

This study makes two unique contributions to the education literature. First, we find that directive coaching causes much more dramatic improvement in participants' teaching in simulations than the more common preparation practice of self-reflection (Hatton & Smith,

1995; Sato, 2014). This is important given that nearly a hundred teacher preparation programs in the United States use this simulation technology to enhance practice experiences, and yet our study is the only one to date to experimentally evaluate the benefits of offering different supports to promote candidates' skill development (Ireland, 2021). Second, this study demonstrates the role that experimental and conceptual replication designs can play in causally identifying methods for supporting teachers, and the conditions in which these supports can be most effective. In our study, we use conceptual replication designs to systematically test hypothesized sources of effect variation. The goal of the conceptual replication effort is to develop foundational knowledge about one potential mechanism–coaching– for supporting teacher candidates that are robust across different populations, pedagogical outcomes, contexts, and settings.

Combined, we view these findings as a critical first step in identifying an evidence-based promising practice -- directive coaching -- to support candidates' development of core instructional skills (Blazar & Kraft, 2015; Hill et al., 2013; Ronfeldt, 2015). We conclude our paper with discussion about ongoing research designed to determine the degree to which these improved practice sessions translate to more distal outcomes of candidates' instructional quality in live classrooms.

Background

Coaching to Support Practice-based Learning in Teacher Education

Preparation programs have long relied on an "apprenticeship" model of clinical practice where candidates learn by observing, practicing, and co-teaching with experienced mentors (Clift & Brady, 2005; Author, 2011). While useful in affording classroom experience, apprenticeship models can be problematic when candidates do not have chances to practice important skills

and/or when mentors model weaker teaching that contradicts principles emphasized in coursework (Feiman-Nemser & Buchman, 1985; Grossman et al., 2009b). Moreover, during clinical placements, candidates often receive feedback about their teaching skills during "triad meetings" with mentors and university-based supervisors that may occur days or weeks after classroom observations (Author, 2011). In contrast, practice in university settings afford scaffolded and uniform opportunities to develop classroom-skills in more controlled and less complex environments, while also receiving immediate feedback from expert teacher educators (Ball & Forzani, 2009; Grossman et al., 2009b).

The traditional model of practice in teacher education also involves attempting larger "chunks" of instruction: a morning meeting, an entire mathematics lesson, a small-group discussion of a text (Clift & Brady, 2005). University-based supervisors are the figures typically conceptualized as the "coach" and work with candidates over the course of a semester or year. Though this form of coaching can be helpful, there is little work, and no causal work to our knowledge that suggests feedback from university supervisors contributes to meaningful changes in candidates' instructional practices (Author, 2011). There are many additional forms of coaching, which necessitate differing levels of time and relational trust between the coach and the individual being coached (Cornett & Knight, 2009; Ericsson & Pool, 2016). Sometimes coaches provide support over extended periods of time (Gibbons & Cobb, 2016). On other occasions, coaches are brought in for shorter durations to target very specific and immediate felt needs (Stahl et al., 2016; Stapleton et al., 2017). There are coaches who focus on the entirety of a professional practice, such as a lead basketball coach. Other coaches concentrate on more specific aspects of practice, such as free throws.

Just as there are multiple forms of coaching, there are multiple forms of practices in which coaching might be integrated. There are occasions when a "full scrimmage" is useful, so that learners can practice coordinating the complex set of skills needed to engage in high-quality classroom instruction (Reich, 2022). There are other occasions when more focused "drills" and aligned coaching help learners "focus on improving more discrete elements" by "abstracting away the complexity of the whole" (Reich, 2022, p. 220). That is, teacher candidates likely need more opportunities to practice and receive coaching on more decomposed aspects of teaching, alongside chances to "recompose" those aspects into entire lessons (Janssen et al., 2015).

When learning a complex practice like teaching, novices often struggle to improve when practice opportunities are longer and more multi-faceted (Van Merriënboer & Kirschner, 2017). In fact, Ericsson and colleagues' (1993) model of deliberate practice hinges on the notion that learners need frequent, low-stakes practice and coaching about focused aspects of the targeted skill. Moreover, practice was shown to be more "deliberate" and useful at leveraging sustained improvements when the coach was clear about the particulars of the desired goal and was explicit in communicating those goals to the learner. Across a wide array of studies, learners benefited from immediate, detailed feedback that compared their performance with those desired goals (Ericsson et al., 1993). We argue that coaching in teacher education could and should also involve more goal clarity around what "good" looks like, coupled with coaching that is tightly aligned with those goals.

The Potential of Simulated Teaching Environments

Digitally mediated simulations, used widely in other professions such as aviation and medicine, offer realistic and standardized practice spaces that can be embedded into coursework, providing a platform to practice, receive coaching, and "try again" (Slater, 2009). Voice actors

(termed “interactors”) who control “student” avatars are trained to respond in real time to candidates’ instructional cues in ways real children would. Importantly, studies have shown that simulations feel more realistic than other approximations like role-plays, and that candidates’ responses are closely aligned with classroom performance (Arora et al., 2011; Dieker et al., 2014).

Simulations are also useful for conducting causal, “basic science” research in applied settings, because they provide a standardized platform for observing candidates and opportunities to systematically vary conditions. Sessions can be delivered in controlled ways, allowing teacher educators and researchers to focus on developing specific skills while limiting other sources of variation, such as instructional content (Author 2014, 2018), the influence of mentors (Goldhaber et al., 2020), or the composition of students (Steinberg & Garrett, 2016). The short duration of simulations also allows candidates opportunities to repeatedly “do-over” teaching scenarios in ways that are impossible in classrooms, while affording real-time coaching that would be logistically challenging during a school day.

Standardized Practice Sessions with Directive Coaching

Because candidates lack the background knowledge and experience to recognize their own strengths and weaknesses, simulated practice alone is likely insufficient to improve candidates’ instructional skills. Candidates also need feedback from experienced coaches who have opportunities to observe the candidate’s practice and can provide concrete, actionable strategies for improvement (Albornoz et al., 2020; Deussen, Coskie, Robinson & Autio, 2007; Hammond & Moore, 2018). This type of directive coaching can also help candidates understand the impact of instruction on students (Author, 2020).

Though one might assume that more coaching is "better," there is no empirical clarity around whether higher dosage coaching is associated with greater observable improvements in instruction (Blazar & Kraft, 2015; Desimone & Pak, 2017; Kraft et al., 2018). In fact, one of the few randomized control trials in this area found no relationship between coaching dosage and teacher outcomes (Pas et al., 2015). Our coaching model – focused on short five-minute sessions with directive feedback – is conceptually similar to many widely used approaches to coaching. For example, "bug in the ear" technology is used in classrooms across the country to provide very short (typically less than 20 seconds) in-the-moment coaching for teachers, and which has been shown to improve teacher outcomes in both the short (Scheeler et al., 2010) and long term (Rock et al., 2014). Recent work by Hanno (2021) provides additional support for the immediate positive associations between short, individual coaching sessions and improvements in teaching. While some of those observed improvements fade-out over time, others do not.

To support candidates' practice and learning in the simulation sessions, we employ a directive, 4-step coaching model where coaches provide targeted feedback on a specific set of instructional skills. The coach first observes the candidate's simulated practice and diagnoses the instructional needs along a skill progression (see example in Appendix A1). Second, the coach gauges the candidate's perception of their performance (e.g., "How are you feeling about the simulation?") before identifying strengths and improvement targets. Third, the coach provides detailed information about the features of high-quality enactment of the targeted skill, how and why it supports positive student outcomes, and specific strategies the candidate can utilize in subsequent simulations. Finally, the coach engages in a role-playing exercise with the candidate, providing opportunities to rehearse a targeted skill. A recent experimental evaluation of this directive coaching model with 100 teacher candidates found large and statistically significant

effects on candidates' observed quality of practice in simulated classroom settings (ES = 1.70 SD), as well as on their perceptions of the student avatars (Author, 2020).

Robustness of Coaching Effects across Key Sources of Variation

While these experimental results suggest that directive coaching *can* improve candidates' teaching, teacher educators also need evidence on the contexts and conditions under which this type of coaching is beneficial (or not) for helping candidates improve. To this end, we identified four theoretically relevant sources of variation in study, participant, and setting characteristics that we hypothesize moderate the magnitude of coaching effects in the MRS. They include:

(1) Timing of study. We want a coaching model that "works" across multiple years, but coaching effects observed at one time may fail to replicate in subsequent years. This may be because effects observed in the first study are the result of statistical chance or error; because the coaching model becomes less effective over time, as coaches begin to deviate from protocols; or because candidates' learning processes change as new technologies and innovations are introduced. Although some coaching studies have looked at impacts across multiple cohorts over time (Blazar & Kraft, 2015; Killion, 2016), these evaluations have largely assessed the impact of changes to coaching models.

(2) Teaching task. Many teachers are not equally skilled across teaching domains (Cohen et al., 2003). A teacher might be strong at establishing warm relationships with students but struggle with providing clear and accurate instructional explanations (Author, 2018; Pianta & Hamre, 2009; Hill et al., 2008). Coaching might also have differential impacts on distinct teaching skills. Kraft and colleagues (2018) find smaller effects for coaching programs focused on content-generic skills (0.07 SD) compared to programs targeting content-specific skills (0.20 SD). However, these results are correlational, and we know comparably little about the benefits

of coaching across different domains of teaching. In our study, we compare the impacts of coaching on two types of pedagogical skills in the MRS: redirecting students' off-task behaviors while "establishing norms" and "supporting text-focused instruction."

(3) Mode of delivery. Although MRS sessions can be delivered remotely over Zoom, we hypothesized that candidates might respond more positively to a coach who observes and supports them in-person. Kraft and colleagues (2018) do not find statistically significant differences between face-to-face and virtual or online coaching but acknowledge they are underpowered to detect meaningful differences. In contrast, Cilliers and colleagues (2020) studied different coaching models for South African teachers and found no effects for virtual coaching, and significant and large effects for in-person coaching. Given that online coaching programs could provide a more resource-effective way of supporting larger numbers of candidates (Israel et al., 2009; Rock et al., 2013; Stapleton et al., 2017), we also compare coaching effects over online Zoom sessions versus in-person sessions.

(4) Target populations and concurrent coursework. Finally, theory suggests that approximations to teaching – like simulations – should not be stand-alone experiences and should be preceded by instruction *about* the approximated teaching practices (Grossman et al., 2009a). Indeed, Kraft et al. (2018) find larger effects of coaching paired with additional training workshops but note the difficulty of disentangling the two because most coaching programs include additional training. In our context, candidates practice and receive coaching feedback on pedagogical skills that are discussed in methods coursework. We theorized that candidates enrolled in a preparation program will be better equipped to utilize and incorporate coaching feedback compared to participants who express interest in becoming teachers but lack formal instruction on pedagogical methods.

Conceptual Replications for Identifying Sources of Variation

In this study, we tested hypothesized sources of effect variation described above through a series of prospectively planned conceptual replication designs (Author, 2019; Author, 2021). The conceptual replication studies were designed according to the Causal Replication Framework, which describes five conditions under which replication studies can be expected to produce the same effects (see Appendix A6). The assumptions may be understood broadly as *replication design* requirements (R1-R2, A6), and *individual study design* requirements (S1-S3, A5). Replication design assumptions include treatment and outcome stability (R1) and equivalence in the causal estimand (R2) across studies. Combined, these two assumptions ensure that the same causal estimand—or a well-defined treatment-control contrast, target population, and setting – is compared across all studies. Individual study design assumptions ensure that valid research designs are used for identifying and estimating unbiased study-specific effects (S1 and S2), and that effects for each study are correctly reported (S3) – they are standard assumptions in most individual causal studies. When one or more of the replication and/or individual study assumptions are not met, replication of effects usually fails.

An implication of the Causal Replication Framework is that it is straight-forward to derive different types of research designs for replication, as well as assumptions required for these designs to yield valid results. *Conceptual replications* examine whether two or more studies with intentionally varied causal estimands yield the same effect. Here, the researcher introduces systematically planned violations in replication assumptions, such as variations in treatment conditions, population characteristics, settings, or outcome measures. Conceptual replication designs include *switching replication designs* with variations in settings across alternating intervention intervals, *multiple cohorts* and *stepped-wedge designs* with variations in

when treatments are introduced across time, and *multi-arm treatment designs* with variations in treatment dosage levels (Author, 2021). In each of these designs, if replication failure is observed, it is because of systematic differences in participants, treatments, outcomes, settings, or time.

Research Methods

To examine the robustness of coaching effects in MRS settings, we use data from five randomized control trials (RCTs) to construct four conceptual replication designs that introduce systematic variations on the dimensions noted above. Figure 1a describes the timing of each of the five individual studies conducted from Spring 2018 to Spring 2020.

Population and Settings

All studies were conducted at a large, selective, public university in the southeast United States. Participants in four of the five experimental studies (Studies 1, 2, 3 and 5) were enrolled in a teacher preparation program that graduates approximately 100 teachers each year. Participants in Study 4 were enrolled in the same university but recruited through an undergraduate course exploring teaching as a profession.[1]

Candidates in Studies 1, 2, 3, and 5 were generally representative of new teachers – that is, they were mostly White, female, and had college-educated parents. The undergraduate sample in Study 4 was less White (60%) and less predominantly female (58%), though also had mostly college-educated parents. Approximately 43% of the undergraduates reported an interest in

[1] All studies included went through the IRB process at the university. Participation in the simulations is a standard part of enrollment in the teacher education program, though students consented to have data included. One of the authors teaches in the preparation program but did not consent students to the research or have access to any data prior to submitting grades for courses.

teaching and 63% reported prior experience working with children (e.g. babysitter, coach; see Table 1 for baseline and setting characteristics for the five studies).

Experimental Design of Individual Studies

For each study, participants were randomly assigned within course sections to receive coaching or to engage with a self-reflection protocol between simulation sessions. Coaches and interactors were scheduled to ensure sufficient variation across course sections, days, and session timings, allowing the research team to control for possible differences due to coaching and interactor effects. Interactors and coaches were shared across study periods, with approximately four coaches per study. Diagnostic results show that for each of the five studies, random assignment was well-implemented. Balance checks of baseline covariates indicate groups were equivalent after randomization (see balance tables for each study Appendices A7-A11). There were no instances of treatment non-compliance where participants failed to "show-up" for coaching sessions or "crossed-over" from the intervention to control conditions (Angrist et al., 1996). For each of the studies, attrition was minimal (less than 15%), with no evidence of differential attrition between groups (see Appendices A7-11 for balance tables and sample sizes for the "full" and "analytic" samples for each study).

Data Collection Procedures

Figure 2 summarizes the data collection procedure for each individual RCT. At Time 1, participants completed questionnaires about their demographic characteristics and teaching experiences, as well as completed baseline simulation sessions where they practiced teaching tasks but did not either receive coaching or self-reflection prompts. Participants were then randomly assigned within course sections to coaching or self-reflection. Approximately two months after baseline sessions, participants completed a second simulation (Time 2).

Immediately after Time 2, half of participants received five minutes of coaching with an expert trained coach, while the other half completed a series of reflection prompts. After the five minutes of coaching or self-reflection, participants completed a third simulation (Time 3), where their pedagogical performance was observed and scored as the outcome.

Simulation Sessions

In each study, participants practiced one of two teaching scenarios. In Studies 1, 3, 4, and 5, participants had to "redirect off-task behaviors" while establishing classroom norms; in Study 2, participants focused on "supporting text-focused instruction" while leading a small-group discussion. For each scenario, participants engaged in a series of "parallel" simulation sessions – meaning that while student avatar responses differed across sessions at Times 1, 2, and 3, they were consistent in terms of the number, type, and intensity of responses. Implementation measures ensured that simulation sessions were delivered consistently across sessions.[2]

Treatment Contrast

Coaching condition. After observing participants at Time 2, coaches provided feedback according to our 4-step protocol. Although feedback focused on different instructional skills and areas of strength and weakness, the structure of the coaching was consistent across studies. Coaches were doctoral students in education who had trained intensively to ensure that the protocol was implemented with fidelity (see footnote 8 for how implementation fidelity of the coaching protocol was assessed).

[2] All actors engage in a multi-week Mursion training, focused on how to use the technological interface. Our research lab engages in a two-day training for all actors and coaches (the two groups are trained separately) on our research scenarios and protocols. To ensure fidelity of implementation, members of our research team then watch a random subset of simulations to provide ongoing feedback and calibration for the actors throughout a research study. Similarly, a member of our research team supports coaches throughout a study, providing ongoing feedback. Coaches and actors are distributed across randomization blocks to maximize variation.

Self-reflection (business-as-usual) condition. Instead of receiving coaching after Time 2, participants in the control condition engaged in a researcher-designed self-reflection protocol that asked participants to identify perceived strengths and weaknesses, and goals for the subsequent simulation session (Yost, 2006). The control condition was consistent across studies.

Measures

To ensure comparability of baseline and outcome measures across studies, we administered the same survey measures in similar settings for each study. Measures were coded using the same protocols and were analyzed in similar ways.

Baseline characteristics. Baseline surveys included information about participants' 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 personality and belief measures including the NEO Five Factor Inventory (McCrae & Costa, 2004), Teacher Sense of Self-Efficacy (Tschannen-Moran & Hoy, 2001), and multi-cultural attitudes surveys (Munroe & Pearson, 2006; see Appendix A4 for a descriptive summary of baseline measures and their psychometric properties).

Outcome for "redirecting off-task behaviors." Our primary outcome measure (Studies 1, 3, 4, and 5) was designed by the research team to align with the Responsive Classroom (2014) framework used across the K-12 schools in which candidates participated in clinical experiences (i.e., student teaching) (for details, see Authors, 2020; Appendix A2 includes the rubric). Responsive Classroom describes an effective redirection as timely, occurring as close as possible to the start of an off-task behavior, thereby minimizing the likelihood of escalation (Kauffman & Landrum, 2006). Effective redirections also specify what a teacher would like a student to do (Good & Brophy, 1995) as succinctly as possible (Levin & Nolan, 2003) to both preserve student

dignity (Sun, 2015) and instructional time (Doyle, 2009). Finally, to preserve a positive learning environment, redirections are delivered calmly, ideally with a positive tone and warm affect (Responsive Classroom, 2014). Simulations were also scored with additional rubrics focused on more discrete aspects of redirection, such as how quickly (in seconds) do participants redirect off-task behavior, or what percentage of the responses provided to student ideas are "perfunctory" (i.e., "good job, "okay"). The overall coaching effects shared here are consistent among these additional outcome measures. Results and rubrics are available upon request.[3]

Scores for the "redirecting off-task behavior" outcome ranged from 1-low to 10-high (Appendices A2).[4] A team of trained and certified raters, blinded to participants' condition, scored videos of all simulation sessions. Fifteen percent of videos were double-scored, with Krippendorff's alpha scores for reliability ranging from 0.75 to 0.88 across studies. Coder drift was addressed with weekly calibration checks and rater agreement reports.

[3] After completing the simulation, all participants completed a short post-simulation survey where they rated student avatars' behavior with a modified IOWA Connor's rating scale (Waschbusch & Willoughby, 2008), a widely-used, brief measure of inattentive-impulsive-overactive (IO) and oppositional-defiant (OD) behavior in children. We find that randomly assigning candidates to coaching yields lower ratings of student avatars as displaying inattentive-impulsive-overactive (IO) and oppositional-defiant (OD) behaviors in Studies 1, 3, and 5, all of which included teacher candidates. However, this shift in ratings was not replicated for Study 4 with undergraduates who were not enrolled in a teacher preparatory program. This suggests that candidates who received coaching evaluate students' behaviors as being less extreme or problematic, which we see as suggestive they are more confident in their skills as managing such behaviors. Just as teacher candidates benefit more from coaching than undergraduates in terms of their observable skill development, they also seem more likely to change their views about children's behaviors than their coached undergraduate counterparts. We see this as important additional evidence of coaching effects on well-known outcomes that are much less tightly aligned with this particular intervention. These analyses are presented in Appendix Table A13.

[4] In exploratory factor analysis on item-level data for each simulation scenario, we found that the optimal factor structure converged on a single factor, suggesting that the teaching task outcome represents a single construct: the focal pedagogical skill for each scenario.

Outcome for "supporting text-focused instruction." We developed the rubrics for the supporting text-focused instruction scenario (Study 2; Appendix A3 includes the rubric) from seminal work on teacher feedback by Hattie and Timperley (2008) and a wealth of reading comprehension research that foregrounds the importance of teacher facilitation of meaningful student interactions with a text (Boardman et al., 2018; Dewitz & Graves, 2021; Duke et al., 2011). In particular, we focus on teachers' text-based questions that: 1) support active engagement with text, 2) help students revise textual misunderstandings, and 3) make text-based arguments (Deshler et al., 2007; Hillocks, 2010; McKeown et al., 2009; Reznitskaya et al., 2009; Shanahan et al., 2010). Teacher feedback during these text-focused interactions is crucial, and Hattie and Timperley (2008) underscore the difference between perfunctory, low-level feedback and descriptive, high-level feedback in which a teacher names the specific, positive features of student contributions – in this case, engaging with a literary text.[5]

[5] We have analyzed the relationship between scores on our "supporting text-focused instruction" rubrics and other, related measures of how teachers engage with student contributions during academic discussions (Author, 2021). These include automated measures of teacher uptake (Demszky et al., 2021) and human-scored measures of how teachers' respond to student ideas, pushing them to clarify and elaborate (Kane et al., 2015). Our measure was significantly correlated with both other measures, providing important convergent validity evidence (Author, 2021). We also have other measurement-focused work underway looking at the relationship between our measures and other measures, including the CLASS (Pianta & Hamre, 2009), though some of this work has been delayed due to shifts to online instruction during the Covid 19 pandemic (Author, 2022). In that work, we elaborate on how, for a number of reasons, the CLASS and other measures like it cannot be used to assess pedagogical quality in simulation sessions. The CLASS is designed to assess broad features of classroom interactions--focused on students and student engagement-- in 15-minute intervals. Our simulations are designed to target more fine-grained, teacher-focused skills and last 5 minutes. Because the CLASS treats the classroom—rather than the teacher—as the unit of analysis, CLASS scores for pre-service teachers are necessarily confounded with characteristics of the student teaching placements in which they work. These include the students in those classrooms, along with the mentor teachers to whom they are assigned. There is certainly value in those measures in those contexts, but we argue there is a distinct and complementary value in the more fine-grained, proximal measures we feature here.

Scores for this outcome measure range ranged from 1-low to 10-high (see Appendices A3 for rubrics). Certified raters scored videos using the same procedures as described above, with Krippendorff's alpha scores for reliability ranging from 0.75 to 0.88.

Effect Variation

To identify *why* heterogeneity in coaching effects may occur, we conducted a series of replication designs that introduced systematic sources of variation across studies, while attempting to ensure that all other study conditions remained the same (Author, 2020, 2021). For ease of interpretation, we selected Study 3 as the "benchmark" study and introduced systematic variations in conceptual replication Studies 1, 2, 4, and 5 for comparing effects. To examine the robustness of coaching effects due to variations in the *timing of the study*, we used a multiple cohort design to compare impacts for candidates from one year (Spring 2018, Study 1) with impacts the following year (Spring 2019, Study 3).

To examine the robustness of coaching effects across *different teaching tasks*, we used a modified switching replication design where participants were randomly assigned to receive coaching at different intervention intervals in alternating sequence such that when one group received coaching, the other group served as the control, and vice versa (Shadish et al., 2002).[6] We compared coaching effects for two intervention intervals, where candidates practiced "supporting a text-focused discussion" (Study 2) in the first period and "redirecting off-task student behaviors" (Study 3) in the second. Here, we assess the replicability of coaching effects across different teaching tasks and pedagogical outcomes with the same sample of participants.

[6] In practice, conditions were rerandomized during the second intervention interval. As a result, some participants were randomized to receive two coaching sessions, one coaching session, or no coaching session across both intervals. There was no evidence of heterogeneity in effects based on the number of coaching sessions received.

To examine the robustness of effects across *different modes of delivery*, coaching effects were compared for in-person simulation and coaching sessions (Study 3) and online through Zoom (Study 5). Finally, to evaluate effects across *different target populations*, we compared results for candidates enrolled in the teacher education program (Study 3) with undergraduates considering careers in teaching but without preparation coursework (Study 4). Figure 1b summarizes sources of variation, replication designs, and study comparisons.

Despite our design-based approach to introduce *systematic* sources of variation across studies, deviations may – and did – occur, where study features differed in more ways than originally intended. For example, participant characteristics may change across studies in a multiple cohort design, fidelity to the coaching protocol may change with in-person versus online delivery, or constructs represented in an outcome measure of quality pedagogical practice may deviate when intervention sessions are focused on different teaching tasks. When study findings are replicated, variations in study populations, contexts, measures, and features provide evidence about the robustness of coaching effects. However, when study findings differ – and there are multiple sources of planned and unplanned variation – it may be difficult for the researcher to determine *why* replication failure occurred (Author, 2018, Author, 2019, Author 2021).

To investigate validity threats to our replication designs – or alternative explanations for why replication failure may have occurred – we present diagnostic information that characterizes the extent to which assumptions were addressed or varied under the Causal Replication Framework. We summarize each study's participant and setting characteristics, fidelity and adherence to the coaching protocol, and study design features (Table 1), along with joint tests of statistical significance (Appendix A5). The diagnostic results presented here are akin to balance statistics commonly reported in experimental designs to demonstrate the comparability (or lack

of comparability) across groups. Appendix A6 summarizes the conclusions of all our diagnostic results evaluating the extent to which replication assumptions were met or violated.

Analysis

To examine the robustness of coaching effects across the five conceptual replication studies, we began by *estimating the conditional average treatment effect* of coaching on participants' pedagogical performance for each individual RCT separately. Coaching effects for each study was estimated using the following model:

$$Y_{ij} = \beta_0 + \beta_1 Coaching_{ij} + \left(X_{ij}\right)\gamma + \delta_i + \alpha_j + \epsilon_{ij} \qquad (1)$$

where, Y_{ij} represents the pedagogical performance for participant *i* in course section *j*, and is a function of participant *i*'s coaching status (where *Coaching*=1 if assigned to receive coaching and *Coaching*=0 if assigned to participate in self-reflection), as well as a vector of characteristics (X_{ij}) measured at baseline, and indicators for any missing baseline information. The model also includes fixed effects for each course section (α_j), which served as blocking factors for random assignment in each study, and for the interactor (δ_i) delivering the simulation session. The coefficient β_1 represents the conditional average treatment effect for each study (see Table 2).

Next, we estimate the *overall average treatment effect* across the five studies using a fixed effects multivariate meta-analytic approach, where each study's effect size was weighted by the inverse variance of the effect estimate.[7] To evaluate *effect heterogeneity* across studies, we

[7] Both the multivariate meta-analytic approach and combined analysis using the pooled data produce very similar results and standard errors. We present the fixed effects meta-analytic result for three reasons. First, the meta-analytic approach provides a straight-forward way to conduct a test of effect homogeneity across replication studies via the Q-test (Hedges & Schauer, 2018). Second, the meta-analytic approach is less restrictive in its modeling assumptions (i.e. it does not assume the same functional form of covariates across the five studies); a fully saturated model using pooled data would be equivalent to our meta-analytic effect. Third, we were able to directly address non-independence in observations for Studies 2 and 3 due to the switching replication design. Here, we adjusted for the dependency structure by first estimating the

examine the *Q*-statistic for the test of homogeneity (Hedges & Schauer, 2018). If the Null hypothesis is rejected and effect heterogeneity is inferred, we compare coaching effects for each set of replication studies to identify the source of the effect variation (see Figure 1b for a summary of study comparisons). If diagnostic results from our balance tables indicate differences in *participant characteristics* across our pairwise comparison of studies, we use propensity score estimation to reweight samples to be observationally similar to our "benchmark" Study 3. We then assess the replicability of "adjusted" coaching effects for observationally similar study samples. If differences in intervention effects persist, we conclude that setting characteristics – as well as unobserved participant characteristics – resulted in treatment effect variation across studies.[8]

Finally, we assessed replication success in study results by comparing the direction, magnitude, and statistical significance patterns of effects, as well as results from a correspondence test that combines statistical tests of difference and equivalence in the same framework (Author, 2018; Tyron, 2001; Tyron & Lewis, 2008). We use the correspondence test because a test of difference may yield ambiguous conclusions when it is underpowered to reject the Null hypothesis of no difference in study effects. To address this concern, the correspondence test combines statistical tests of difference and equivalence into the same framework to distinguish between results that are: statistically different (significant difference, non-significant

correlation in effect estimates using microdata for Studies 2 and 3 and a bootstrapping procedure, and then by adjusting the covariance-variance matrix in the meta-analytic effect that accounts for the estimated correlation in effects. The multivariate meta-analysis was conducted using Metafor (Viechtbauer, 2010), which allows users to specify the covariance-variance matrix for effect estimates included in the meta-analysis.

[8] Details of the propensity score estimation procedure for creating weights, balance characteristics for study sample comparisons, and estimation of the adjusted coaching effects appear in Appendix A12.

equivalence), statistically equivalent (non-significant difference, significant equivalence), statistically indeterminate (non-significant difference *and* equivalence), or trivial difference (significant difference *and* equivalence).[9] We report conclusions using a stringent threshold (0.2 SD) for equivalence (Author, 2022), and a more generous threshold (1 SD), given prior evidence that directive coaching effects are larger than 1 SD (Author, 2018).

Results

Diagnostic Results of Replication Assumptions

Variations in participant and setting characteristics. Table 1 summarizes participant and setting characteristics for each study to demonstrate the extent to which these factors systematically varied and/or replicated across studies. The bolded text highlights systematic differences in participant and setting characteristics introduced across replication efforts. Appendix A5 displays the effect size difference in study sample characteristics for each conceptual replication, as well as results of joint tests of statistical significance comparing differences in characteristics across sets of studies. We find that for Studies 1, 2 and 3, teacher candidate samples were qualitatively similar in terms of demographic characteristics and contextual experiences (Table 1). Although our balance statistics and F-test results indicate differences between samples (Appendix A5), they were due to a few participants who indicated variations from an overall homogeneous sample that consisted of mostly white women with college-educated parents who attended suburban high schools.

Participant and setting characteristics for Study 4, however, deviated substantially from the other three studies. Participants in Study 4 had different undergraduate course experiences,

[9] The Correspondence Test was implemented through a STATA program written by Author (2023) based on results presented in Author (2022).

were younger, more male, and less white than the participants (teacher candidates) in other studies. Since they were enrolled in an undergraduate course that explored teaching as a profession (and not in the teacher preparation program), they also lacked prior course experiences on pedagogical methods. The deviations reflect differences in the underlying target populations and contexts from which the research team sought to sample. Participants in Study 5 were also different from those represented in Studies 1, 2, and 3 – they were younger, less white and more likely to have attended an urban high school or a high school with majority high-achieving students. This compositional shift reflects a policy change for the preparation program between the two periods. In Studies 1, 2, and 3, the preparation program included a combined BA/Master's degree program, while in Study 5, the combined degree program was separated into distinct Bachelor's and Master's degree programs. While students in both degree programs underwent a similar sequence of course work, undergraduate and graduate students had different participant characteristics.

Variations in coaching delivery. To examine the extent to which the coaching intervention was delivered consistently across studies, the research team used a natural language processing method to quantify the semantic similarity between a benchmark coaching protocol and transcripts of coaching sessions as delivered. Author (2021) demonstrates that semantic similarity scores can be used to assess intervention fidelity in evaluation settings with highly standardized protocols that are delivered through verbal interactions with participants.[10]

[10] In these cases, texts from transcripts of intervention sessions can be represented by their vocabulary and compared to one another by the relative frequency with which they use a set of words or phrases. In our context, we use semantic similarity methods to quantify adherence to the coaching protocol by comparing coaching transcripts with scripted protocols that the research team has identified as gold standard "benchmarks" for high-quality coaching delivery. See Author (2021) for further description of the method and interpretation of scores.

Adherence scale scores obtained from the method range from 0 to 1, with transcripts of coaching sessions with high adherence to the coaching protocol having higher scores, and those that stray from the protocol having lower scores. Adherence scores in Table 1 indicate that fidelity to the coaching protocol was similar across studies, though coaching fidelity was higher in Study 2 (0.38) relative to the other studies, which ranged in scores from (0.20-0.26).

Summary. Combined, these results suggest that while the conceptual replication studies succeeded in introducing systematic variations in time (Study 1 vs. Study 3), teaching task and outcome (Study 2 vs Study 3), participant and setting characteristics (Study 3 vs. Study 4), and online versus in-person delivery (Study 3 vs Study 5), there were also unplanned deviations from the original replication designs. To investigate the role of participant characteristics in moderating coaching effects across studies, we will examine the replicability of results for samples that have been adjusted to appear observationally similar on participant characteristics.

Impact of Coaching on Participants' Instructional Practices

Table 2 presents effect size estimates of coaching on the quality of participants' pedagogical practices in the simulations. Columns 1-5 in Table 2 provide separate effect estimates in standard deviation units for each study. Effect sizes ranged from 0.39 SD (p-value = *ns*; Study 4) to 1.69 SD (p-value < 0.01; Study 1). Coaching effects for candidate samples (Studies 1, 2, 3, and 5) were consistently large and statistically significant (ranging from 1.41 SD in Study 2 and 3 to 1.69 SD in Study 1), while the coaching effect for undergraduates (Study 4) was 0.39 SD and not statistically significant. Across the five studies, the multivariate meta-analytic coaching effect was positive, large, and statistically significant (1.34 SD, p-value <

0.01).[11] However, the test of homogeneity indicated significant differences in effect estimates across studies (Q-statistic = 21.99; df = 4; p-value < 0.01).

Robustness of Results Across Systematic Sources of Effect Heterogeneity

Given evidence of effect heterogeneity, we also examined results from our replication studies to identify sources of effect variation. Table 3 summarizes results from each set of replication comparisons (with unadjusted and adjusted effects), with ✓ indicating replication success by a pre-specific criterion (magnitude, sign, statistical significance pattern of results), and X indicating replication failure by the pre-specified criterion. We also report the estimated difference in effects and whether they are statistically different, and conclusions from the correspondence test (difference, equivalence, indeterminant, trivial difference).

Timing of study. Results from Table 3 show that coaching effects were robust across variation in study timing under most criteria for replication success. The coaching effect for Study 3 was 1.41 SD (p-value < 0.01), while the coaching effect was 1.69 SD (p-value < 0.01) for Study 1. When the Study 1 sample was reweighted to be observationally similar to participants in the benchmark Study 3, the adjusted coaching effect for Study 1 was 1.45 SD. Overall, we find that effects were replicated in terms of direction, magnitude, and statistical significance patterns. The difference in effect estimates was also not statistically significant. In looking at results from the correspondence test, we find that effects are statistically indeterminant

[11] The combined analysis with the pooled data yielded a similar result, with an overall treatment effect of 1.35 SDs (0.053).

(neither statistically different nor equivalent) with the more stringent tolerance threshold (0.2 SD), but statistically equivalent with the empirically-based threshold of 1 SD.[12]

*Teaching task***.** When participants practiced "establishing classroom norms," coaching improved their performance by 1.41 SDs (p-value <0.01, Study 3); when they practiced "text-focused instruction," coaching improved performance by 1.41 SDs (p-value < 0.01, Study 2) or by 1.42 SDs (p-value < 0.01, Study 2) for the adjusted coaching effect. Coaching effects were comparable in terms of magnitude, direction, and statistical significance patterns, and the correspondence test again indicated indeterminacy for the equivalence test at the 0.2 SD threshold but equivalence for the 1 SD threshold.

Mode of delivery. The coaching effect for the face-to-face sessions was 1.41 SD (p-value <0.01, Study 3). The unadjusted coaching effect for sessions delivered on Zoom was 1.62 SD (p-value<0.01, Study 5), and the adjusted coaching effect was 1.37 SDs. These results indicate that while reweighting the Study 5 sample to be more observationally similar to the Study 3 sample produced an effect closer to the benchmark study result, coaching effects for both the weighted and unweighted samples are comparable. Across most criteria – direction, magnitude, statistical significance, and difference – replication success was achieved. Although the correspondence test yields a conclusion of statistical indeterminacy for a tolerance threshold of 0.2 SD, it concludes that effects are equivalent for the larger tolerance threshold of 1 SD.

Target Population and Concurrent Coursework. Finally, for candidates who were enrolled in methods classes, the coaching effect was 1.41 SDs (p-value <0.01, Study 3), but for undergraduates not enrolled in preparatory courses, the effect was smaller and not statistically

[12] As highlighted by Hedges & Schauer (2018) and Author (2022), the conclusion of "statistical indeterminancy" for a stringent tolerance threshold of 0.2 SD highlight the challenge of low statistical power for many pairwise replication studies.

significant (0.39 SDs; Study 4). When the participant sample in Study 4 is reweighted to appear similar to the Study 3 sample, the adjusted coaching effect is 0.67 SDs. Combined, these results suggest that differences in observable participant characteristics explained some but not all of the variation in coaching effects across Studies 3 and 4. Instead, coaching effects were likely also moderated by contextual characteristics (including concurrent course experiences) and other unobserved participants characteristics. For both the unadjusted and adjusted effects, we conclude replication failure of results in terms of magnitude of effect and statistical significance pattern. The correspondence test yielded a conclusion of statistical difference for the more stringent 0.20 SD tolerance threshold and for the larger 1 SD threshold (significant difference, non-significant equivalence).

Discussion and Implications

Teacher preparation needs more evidence, particularly causal evidence, about promising practices for expediting teacher learning and skill development. Coaching – used extensively with practicing teachers – has been shown to improve a range of outcomes from instructional skills to student achievement (Allen et al., 2011; Cilliers et al., 2020; Kraft et al., 2018). Our early work in a pre-service context suggests that targeted, individualized, and directive coaching can also improve candidates' instructional skills (Authors, 2020). Given the resource intensive nature of coaching, however, we need more causal evidence about the robustness of coaching effects, as well as the contexts and conditions under which coaching is likely to be effective.

Here we use conceptual replication research designs to implement five RCTs that evaluate the impact of directive coaching, using simulated classrooms to both approximate and assess teaching. Across four studies, we see significant performance improvements because of coaching. This provides encouraging evidence that teacher preparation can be an important time

for rapid skill development, when candidates are given targeted practice opportunities and corresponding supports. Though we often think that practice has to happen in real classrooms with real children, we provide robust evidence that "the work of teaching" can be incorporated into coursework (Ball & Forzani, 2009). Rather than waiting until candidates are in clinical placements, providing structured practice and targeted feedback in ways that are integrated across coursework can better prepare candidates for skills with which they often report struggling (Grossman et al., 2009a).

We also find that directive coaching can leverage large improvements, even absent long standing relationships between candidates and coaches. Though many have argued for the value of responsive coaching, where coaches cultivate trust with the teachers they support, we find robust evidence that coaches who do not know candidates – and support them in only brief, directive, skill-focused sessions – can promote rapid skill developments (Killion, 2016; Steiner & Kowal, 2007). This is not to argue that relationships are not important in teacher education, but our data suggest that additional, less time-intensive supports also can be effectively layered onto practice experiences. We recognize that our model of coaching where the candidate and the coach are working towards clearly defined and articulated goals is not the norm. That said, our data suggest that teachers, like all learners, can benefit from a clear understanding of what they are working towards and why (Ericsson & Pool, 2016). These cycles of repeated practice with coaching around tangible and well-articulated goals are the norm in many professional fields, and we see no reason why the same could not be true in teacher education (Reich, 2022).

This study is the first to our knowledge that uses a series of systematic replication studies to inform theory about how, when, and for whom coaching "works" in a field where we have next to no causal evidence. Since each study was designed prospectively, the research team

introduced systematic sources of variation to examine heterogeneity in observed coaching effects across different teaching tasks, timing of study, targeted participants, and modes of coaching (Authors, 2021). Our findings suggest that coaching significantly improves candidates' teaching skills, and that coaching effects replicate across pedagogical tasks, timing, and modes of delivery. This is encouraging for programs looking for ways to integrate simulations and coaching (Dieker et al., 2014).

We also find that coaching effects are not robust across participants and contextual experiences. Undergraduates did not improve from coaching as much as candidates enrolled in concurrent methods coursework focused on the practices targeted in simulations. Though our data do not allow for definitive conclusions about mechanisms, we theorize smaller coaching effects for the undergraduate sample, even after controlling for observable characteristics, may be explained by their lack of schema or prior knowledge about the skills targeted in coaching. This suggests that coaching in isolation, without corresponding coursework on targeted practices, is not as effective (Kraft et al., 2018). It also underscores the importance of coherent and coordinated learning experiences where candidates engage with the theory underlying teaching practices, have opportunities to observe and analyze use of such practices, and *then* have chances to enact those practices with coaching supports (Grossman et al., 2009a). That is, approximations of teaching should not be stand-alone experiences, where skills are decoupled from their conceptual bases (Kennedy, 2016). This is in line with previous studies that highlight the importance of grounding in-service coaching with corresponding instruction about related skills (Albornoz et al., 2020; Kraft et al., 2018; Scheeler et al., 2009). Pre-service coaching programs might want to develop cycles of learning that ensure skills practiced and coached build on a

robust foundation of knowledge about the skills, what they look like in use in classrooms, and how and why they support positive student outcomes.

There are, however, limitations to these study results, particularly in that we are not able to observe the longer-term effects of our coaching model on less tightly-aligned measures of teaching quality in K-12 classrooms. That said, the primary purpose of this work has been focused on causally identifying supports for improving teacher practice in simulated settings, *before* we invest the much more substantial resources in tracking the development of those skills in more applied contexts. Nearly a hundred teacher preparation programs are using the simulation technology we rely on here. And yet, prior to our work, the field knew next to nothing about the utility of different approaches to simulated practice. The vast majority of programs were asking candidates to practice without additional supports (Ireland, 2021), assuming the accuracy of the adage "practice makes perfect." Many other programs were asking candidates to self-reflect on their performance in the simulated classroom (Ireland, 2021). This focus on "self-reflection" as a lever for improvement is a central tenet of much teacher education practice and policy, as demonstrated by the fact that one of the most common licensure exams, edTPA, primarily assesses prospective teachers' reflection skills (Sato, 2014).

Thus, our central goal across the studies presented here was understanding if coaching helped teachers improve more in simulated contexts than other less-resource intensive approaches to simulation, as well as the contexts and conditions in which coaching helped more or less. In other words, we wanted to understand how to get simulated practice "right" first, before looking at the robustness and transferability of our methods.

The effect sizes presented here are large – much larger than what is typically observed in educational research generally (Kraft, 2020; Lipsey et al., 2012), and in the study of coaching

interventions specifically (Kraft et al., 2018). We acknowledge that our large effect sizes are likely bolstered by the proximal and aligned nature of the outcome measures, though we also present evidence of coaching effects on less tightly aligned measures like the IOWA Connor's rating scale. It will be important to analyze the degree to which our observed coaching effects persist across the teacher preparation period or fade-out over time. We are in the midst of this work, though it has been impacted by shifts in teacher preparation during the Covid-19 pandemic (Author, 2022). We also need to build more robust evidence about correspondence between improved teaching in simulated classrooms and improvements in the more distal outcomes of teaching real children in real classrooms. This work is also currently underway (Author, 2021). Finally, all of our work has been done at a single university with a specific population of candidates. 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 sites, samples, and teaching outcomes is a critical area for ongoing and future work. Despite these limitations, we see these results as an important first step in identifying simulations and directive coaching as an efficient and effective method for helping candidates develop important teaching skills. Given the short duration and crucial importance of teacher preparation, building a robust evidence base about such methods is imperative.

References

Albornoz, F., Anauati, M. V., Furman, M., Luzuriaga, M., Podestá, M. E., & Taylor, I. (2020). Training to teach science: experimental evidence from Argentina. *The World Bank Economic Review*, *34*(2), 393-417.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034-1037.

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

Arora, S., Miskovic, D., Hull, L., Moorthy, K., Aggarwal, R., Johannsson, H., ... & 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.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 2332858415607834.

Ball, D. & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education*, *60*(5), 497-511.

Blazar, D. & Kraft, M. A. (2015). Exploring mechanisms of effective teacher coaching: A tale of two cohorts from a randomized experiment. *Educational Evaluation and Policy Analysis*, *37*(4), 542-566.

Boardman, A. G., Boelé, A. L., & Klingner, J. K. (2018). Strategy instruction shifts teacher and student interactions during text-based discussions. *Reading Research Quarterly*, *53*(2), 175-195.

Bowman, C. L., & McCormick, S. (2000). Comparison of peer coaching versus traditional supervision effects. *Journal of Educational Research*, *93*(4), 256-261.

Castro, J. F., Glewwe, P., Heredia-Mayo, A., & Montero, R. (2021). *Work with What You've Got: Improving Teachers' Pedagogical Skills at Scale in Rural Peru* (No. 1701-2021-3436).

Cilliers, J., B., Fleisch. C. Prinsloo, and S. Taylor (2020). How to improve teaching practice? Experimental comparison of centralized training and in-classroom coaching. *Journal of Human Resources 55*(3), 926– 962.

Clift, R. T., & Brady, P. (2005). Research on methods courses and field experiences. *Studying teacher education: The report of the AERA panel on research and teacher education*, *309424*.

Coburn, C. E., & Woulfin, S. L. (2012). Reading coaches and the relationship between policy and practice. *Reading Research Quarterly*, *47*(1), 5-30.

Cronbach, L. J. (1982). In praise of uncertainty. *New Directions for Program Evaluation*, *1982*(15), 49–58. https://doi.org/10.1002/ev.1310

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2021). *Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course.* EdWorkingPaper No. 21-483. *Annenberg Institute for School Reform at Brown University*.

Deshler, D. D., Palincsar, A. S., Biancarosa, G., & Nair, M. (2007). Informed choices for struggling adolescent readers: A research-based guide to instructional programs and practices. *International Reading Association (NJ3)*.

Desimone, L. M., & Pask, K. (2017). Instructional coaching as high-quality professional development. *Theory Into Practice*, *56*(1), 3-12.

Dewitz, P., & Graves, M. F. (2021). The science of reading: Four forces that modified, distorted, or ignored the research finding on reading comprehension. *Reading Research Quarterly*, *56*, S131-S144.

Dieker, L. A., Rodriguez, J. A., Lignugaris/Kraft, B., Hynes, M. C., & Hughes, C. E. (2014). The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, *37*(1), 21-33.

Duke, N. K., Pearson, P. D., Strachan, S. L., & Billman, A. K. (2011). Essential elements of fostering and teaching reading comprehension. *What research has to say about reading instruction*, *4*, 286-314.

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*.

Doyle, W. (2009). Situated practice: A reflection on person-centered classroom management. *Theory Into Practice, 48*(2), 156-159.

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

Feiman-Nemser, S., & Buchmann, M. (1985). Pitfalls of experience in teacher preparation. *Teachers College Record*, *87*(1), 53-65.

Goldhaber, D., Krieg, J., Naito, N., & Theobald, R. (2020). Making the most of student teaching: The importance of mentors and scope for change. *Education Finance and Policy*, *15*(3), 581-591.

Good, T. L. & Brophy, J. E. (1995). *Contemporary educational psychology*. Longman/Addison Wesley Longman.

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: theory and practice*, *15*(2), 273-289.

Hammond, L. & Moore, W. M. (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.

Hanno, E. C. (2021). Immediate changes, trade-offs, and fade-out in high-quality teacher practices during coaching. *Educational Researcher*, 0013189X211062896.

Harris, D. N. & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, *95*(7-8), 798-812.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Hedges, L. V. & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265-275.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education*, *39*(4), 372-400.

Hill, H. C., Beisiegel, M., & Jacob, R. (2013). Professional development research: Consensus, crossroads, and challenges. *Educational Researcher*, *42*(9), 476-487.

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 pre-service teachers? A literature review. *Teaching and Teacher Education*, *52*, 99-112.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499-534.

Ireland, A. (2021). *Mixed reality simulation in teacher preparation programs in the United States.* [Unpublished doctoral dissertation]. University of California, Los Angeles.

Israel, M., Knowlton, H. 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.

Janssen, F., Grossman, P., & Westbroek, H. (2015). Facilitating decomposition and recomposition in practice-based teacher education: The power of modularity. *Teaching and Teacher Education*, *51*, 137-146.

Kane, T., Hill, H., & Staiger, D. (2015). National Center for Teacher Effectiveness Main Study. (ICPSR36095-v2).

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.

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

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.

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.

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*, *33*(4), 279-299.

Landrum, T. J. & Kauffman, J. M. (2006). Behavioral approaches to classroom management. *Handbook of classroom management: Research, practice, and contemporary issues*, 47-71.

Levin, J. & Nolan, J. F. (2003). *What every teacher should know about classroom management*. Pearson Education.

Majerowicz, S. & Montero, R. (2018). *Can Teaching be Taught? Experimental Evidence from a Teacher Coaching Program in Peru* (Working paper).

Matsko, K. K., Ronfeldt, M., Nolan, H. G., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*, *71*(1), 41-62.

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.

McKeown, M. G., Beck, I. L., & Blake, R. G. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, *44*(3), 218-253.

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.

Papay, J. 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.*

Pas, E. T., Bradshaw, C. P., Becker, K. D., Domitrovich, C., Berg, J., Musci, R., & Ialongo, N. S. (2015). Identifying patterns of coaching to support the implementation of the Good Behavior Game: The role of teacher characteristics. *School Mental Health*, *7*(1), 61-73.

Pianta, R. C. & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, *38*(2), 109-119.

Reich, J. (2022). Teaching drills: Advancing practice-based teacher education through short, low-stakes, high-frequency practice. *Journal of Technology and Teacher Education*, *30*(2), 217-228.

Reisman, A., Cipparone, P., Jay, L., Monte-Sano, C., Kavanagh, S. 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.

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

Reznitskaya, A., Kuo, L. J., Clark, A. M., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education*, *39*(1), 29-48.

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.

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, *66*(4), 304-320.

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education*, *65*(5), 421-434.

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.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experiments and generalized causal inference. *Experimental and quasi-experimental designs for generalized causal inference*, 1-32.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). Improving Reading Comprehension in Kindergarten through 3rd Grade: IES Practice Guide. NCEE 2010-4038. *What Works Clearinghouse*.

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.

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

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.

Steinberg, M. P. & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure?. *Educational Evaluation and Policy Analysis*, *38*(2), 293-317.

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*.

Sun, R. C. (2015). Teachers' experiences of effective strategies for managing classroom misbehavior in Hong Kong. *Teaching and Teacher Education*, *46*, 94-103.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6, 371–386.

Tryon, W. W., Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods*, 13, 272–278.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783-805.

Van Merriënboer, J. J. G. & Kirschner, P. A. (2017). *Ten steps to complex learning: A systematic approach to four-component instructional design (3rd. Edition)*. Routledge.

Viechtbauer, W (2010). Conducting meta-analysis in R with metaphor package, *Journal of Statistical Software, 36*(3), 1-48.

Yost, D. S. (2006). Reflection and self-efficacy: Enhancing the retention of qualified teachers from a teacher education perspective. *Teacher Education Quarterly*, *33*(4), 59-76.

Figure 1a
Planned Replication Studies from 2017 through 2020

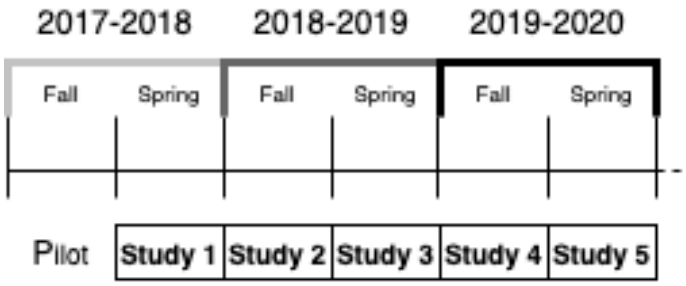
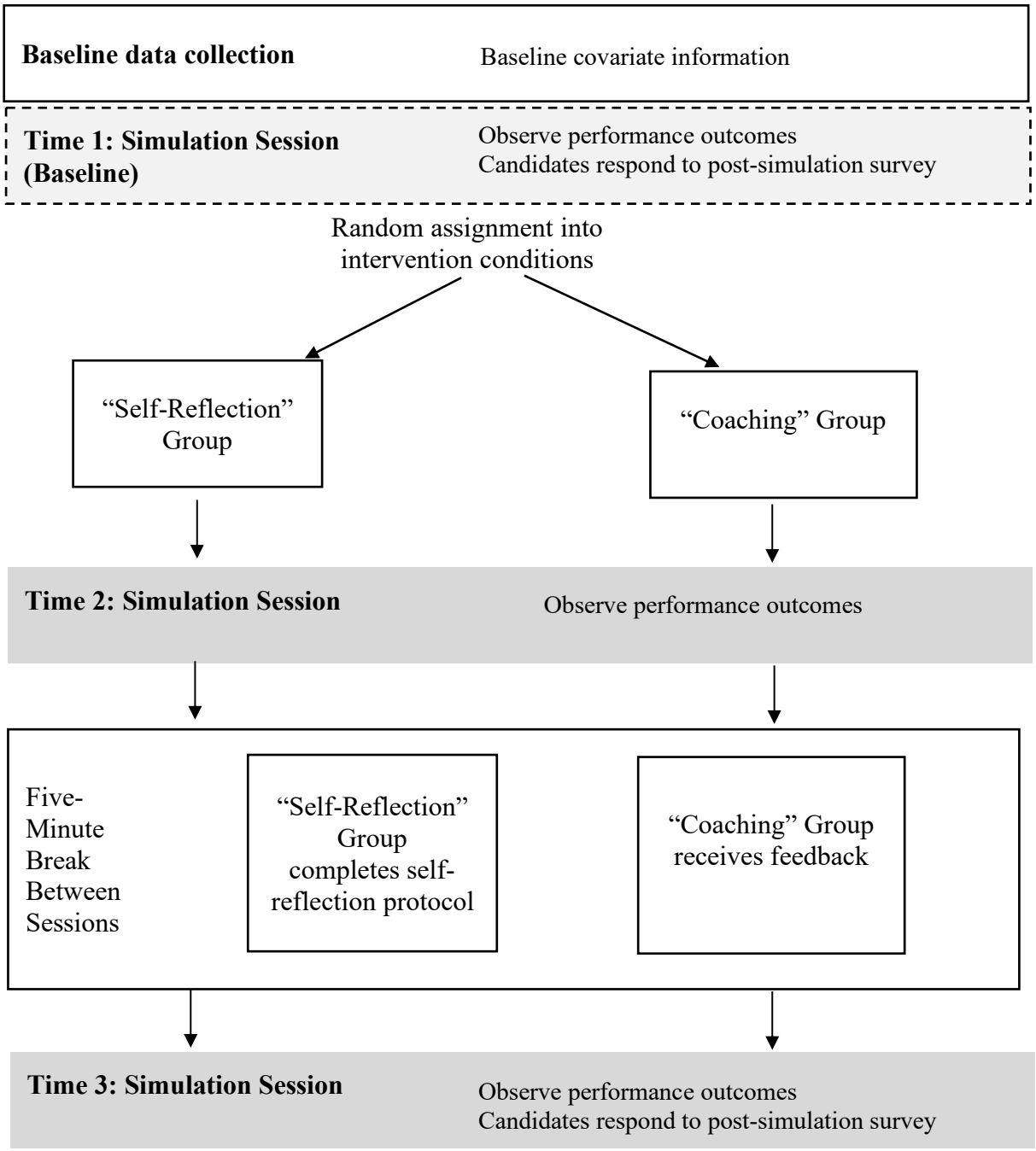


Figure 1b
Conceptual Replication Designs for Understanding Sources of Systematic Variation

Source of Variation	Replication Design	Study Comparison
Timing of study (Spring 2019 vs. Spring 2018)	Multiple cohort design	Study 3 vs. Study 1
Teaching task (Establishing norms vs. Supporting text-focused instruction)	Switching replication design	Study 3 vs. Study 2
Mode of delivery for practicing and coaching sessions (In-person vs. Online)	Conceptual replication design	Study 3 vs. Study 5
Participant characteristics and concurrent coursework (Teacher candidates vs. undergraduates interested in teaching)	Conceptual replication design	Study 3 vs. Study 4

Notes: For ease of interpretation, we selected Study 3 as the "benchmark study" for comparing study effects across the different systematic replication designs. Conceptual replication designs are based on research designs introduced by (Author, 2021; Author, 2019). See Appendix A6 for description of conceptual replication design and their assumptions.

Figure 2
Data Collection Procedure for Individual RCT Studies



Notes: The data collection protocol for Study 4 deviated slightly from the other three studies depicted in Figure 2. Because Study 4 was conducted as part of an undergraduate class, participants completed assessments of their demographic characteristics and experiences at Time 1 but did not engage with the baseline simulation session. For Study 4, performance on the teaching task at Time 2 provided pre-intervention scores of the outcome, and performance at Time 3 provided quality of pedagogical instruction scores.

Table 1

Descriptive Statistics across Replication Studies

	Study 1 (Spring 2018)	Study 2 (Fall 2018)	*Benchmark Study 3 (Spring 2019)*	Study 4 (Fall 2019)	Study 5 (Spring 2020)
Participant Characteristics					
GPA	3.44	3.48	3.43	3.49	3.52
% Either parent a teacher	0.23	0.35	0.36	0.27	0.22
% Mother education- college or above	0.83	0.96	0.96	0.87	0.82
% Father education- college or above	0.69	0.83	0.83	0.87	0.76
% Female	0.74	0.71	**0.69**	**0.58**	0.68
% Over the age of 21	0.95	0.83	0.86	0.55	0.78
% White	0.80	0.85	**0.86**	**0.60**	0.73
Location of high school attended					
% Rural	0.22	0.24	0.26	0.06	0.12
% Suburban	0.74	0.67	0.75	0.84	0.35
% Urban	0.05	0.11	0.01	0.10	0.48
Average SES of high school attended					
% Low SES	0.06	0.05	0.05	0.07	0.01
% Middle SES	0.62	0.79	0.77	0.61	0.24
% High SES	0.35	0.22	0.23	0.32	0.51
Majority race of high school attended					
% Primarily students of color	0.07	0.08	0.07	0.10	0.05
% Mixed	0.43	0.48	0.46	0.35	0.19
% Primarily white students	0.54	0.53	0.56	0.55	0.41
Average achievement level of high school attended					

% Primarily low achieving	0.08	0.04	0.04	0.06	0.05
% Primarily middle achieving	0.45	0.61	0.57	0.39	0.46
% Primarily high achieving	0.47	0.35	0.39	0.55	0.49
Instructional quality performance score at pretest	3.64	3.94	**3.46**	**2.87**	2.82
Setting Characteristics					
Timing	**Spring 2018**	Fall 2018	**Spring 2019**	Fall 2019	Spring 2020
Teaching task	Establishing classroom norms	**Supporting Text-Focused Instruction**	**Establishing classroom norms**	Establishing classroom norms	Establishing classroom norms
Mode of delivery	In-person	In-person	**In-person**	Online	**Online**
Participant Characteristics and Concurrent Coursework	Teacher Preparation (Methods Course)	Teacher Preparation (Methods Course)	**Teacher Preparation (Methods Course)**	**Undergraduate Program (Teaching as a Profession Course)**	Teacher Preparation (Methods Course)
Study Characteristics					
Adherence to Coaching Model Delivery	0.23	0.38	0.25	0.22	0.20
Research Design (Assignment to Coaching)	RCT	RCT	RCT	RCT	RCT
Initial sample N	105	119	117	115	113
Full sample N	102	111	98	99	112

Notes: Demographic information comes from data collected by the teacher preparation program or administered as surveys to study participants. Each row represents regression-adjusted means for each study from a separate regression with the same right-hand specification but different covariates as the dependent variable. Models include controls for randomization blocks. Bold text highlight study characteristics that were planned sources of variation across individual studies (using Study 3 as the "benchmark study" for

comparing study effects from 1, 2, 4, and 5). “Adherence to Coaching Model Delivery” was assessed using the semantic similarity approach described in Author (2021); a higher score indicates higher similarity to a benchmark scripted treatment protocol. To examine the validity of the RCT, the research team examined equivalence on an array of baseline characteristics for each study. See Appendices A7-A11 for balance tables of individual studies. The “initial sample” includes all participants in each study who were randomly assigned into either the coaching or self-reflection conditions. The “full sample” includes participants in each study who were randomly assigned and completed baseline measures.

Table 2
Coaching Effect Sizes by Study

	Study 1	Study 2	Study 3	Study 4	Study 5
	(1)	(2)	(3)	(4)	(5)
Providing Text-Based Feedback (SD)		1.41**			
		(0.21)			
Redirecting Off-Task Behaviors (SD)	1.69**		1.41**	0.39	1.62**
	(0.22)		(0.21)	(0.23)	(0.19)
Constant	-1.68	-0.80	-0.61	-0.38	-1.09
	(0.54)	(0.54)	(0.37)	(0.30)	(0.28)
Analytic sample N	99	99	90	95	102

Notes: All adjusted coaching effects are presented in effect sizes. Coefficients and standard errors (in parentheses), and constant values are reported in columns (1) through (5). Models for each study-specific effect include controls for randomization blocks, participants' gender, race, high school GPA, baseline score, indicators for missing baseline values, and interactor fixed-effects. In specification checks of individual study effects, we found no evidence that coaching effects varied by course sections (blocking factor) or by interactor. The "analytic sample" includes participants who were randomized, completed baseline and post-test measures on the outcome. Across the five studies, the multivariate meta-analytic effect is 1.34 SDs (0.094) and the overall effect from the pooled data is 1.35 (0.053). +p < .10. *p < .05. **p < .01

Table 3

Replication Success across Series of Systematic Replication Studies

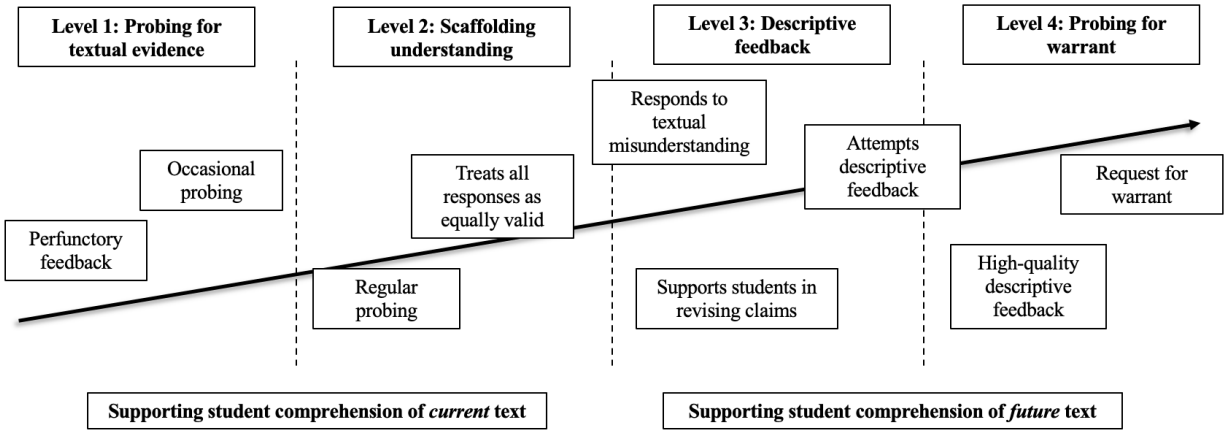
Source of Variation	Studies	Treatment effect	Magnitude of effects	Sign of effects	Significance patterns	Estimated difference between coaching effects	Outcome of Correspondence Test (Δ= 0.2 SD)	Outcome of Correspondence Test (Δ= 1 SD)
Timing of study (Spring 2019 vs. Spring 2018)	Study 3 (N=90) *vs.* Study 1 (N=99)	1.41** (0.21) 1.69** (0.22)	✓	✓	✓	-0.28 (0.31)	Indeterminacy	Equivalence
	Adjusted Study 3 (N=90) *vs.* *Adjusted* Study 1 (N=99)	1.41** (0.21) 1.45** (0.33)	✓	✓	✓	-0.03 (0.39)	Indeterminacy	Equivalence
Teaching task (Establishing norms vs. Supporting text-focused instruction)	Study 3 (N=90) *vs.* Study 2 (N=99)	1.41** (0.21) 1.41** (0.21)	✓	✓	✓	-0.00 (0.33)[#]	Indeterminacy	Equivalence
	Adjusted Study 3 (N=90) *vs.* *Adjusted* Study 2 (N=99)	1.41** (0.21) 1.42** (0.21)	✓	✓	✓	-0.01 (0.33)[#]	Indeterminacy	Equivalence
Delivery (In-person vs. Online)	Study 3 (N=90) *vs.* Study 5 (N=102)	1.41** (0.21) 1.62** (0.19)	✓	✓	✓	-0.21 (0.28)	Indeterminacy	Equivalence

	Adjusted Study 3 (N=90) *vs.* *Adjusted* Study 5 (N=102)	1.41** (0.21) 1.37** (0.23)	✓	✓	✓	-0.04 (0.32)	Indeterminacy	Equivalence
Participant characteristics and concurrent coursework (Teacher candidates vs. undergraduates interested in teaching)	Study 3 (N=90) *vs.* Study 4 (N=95)	1.41** (0.21) 0.39 (0.23)	×	✓	×	-1.02** (0.31)	Difference	Difference
	Adjusted Study 3 (N=90) *vs.* *Adjusted* Study 4 (N=95)	1.41** (0.21) 0.67** (0.26)	×	✓	✓	-0.74** (0.33)	Difference	Difference

Notes: For each criterion for replication success, ✓ indicates that replication success was achieved, × indicates that replication failure occurred. Correspondence test results (Author, 2018; Tyron, 2001; Tyron & Lewis, 2008) indicate whether we have evidence to conclude that the two study effect estimates are statistically different, statistically equivalent, statistically indeterminant (not statistically different or equivalent), or trivially different (both statistically different and equivalent). For the unadjusted effects, correspondence test results were conducted using coefficients and standard errors obtained from Table 2. For the adjusted effects, correspondence test results were conducted based on coefficient effects and standard errors obtained from the matched samples. [#] We use a bootstrapping procedure described by Author (2018) to calculate the standard error for the difference test in the modified switching replication design. The bootstrapped standard error accounts for non-independence in study effects due to shared participants in Studies 3 and 2. *Adjusted* study effects includes treatment effect estimates obtained from matching sample participants in Studies 3 and the respective study (Study 1, 2, 4, and 5) on baseline demographic characteristics using inverse propensity score weights; see Appendix A12 for technical details on how matching was performed. The Ns represent the "analytic sample" who were randomized in each study and completed baseline and post-intervention measures on the outcome. +p < .10. *p < .05. **p < .01

Appendix A

A1. Example Skill Progression for the Supporting Text-Focused Instruction Teaching Scenario



Examples: • "Nice job!" • "Interesting." • "Hmmm. I like that!" • "That works!"	Examples: • "What makes you think that?" • "How do you know?" • "Tell me more."	Examples: • "Great. I got something different but that's ok." • "Readers interpret the text differently, so that works for me."	Examples: • "What in the story made you think that?" • "Let's read paragraph 19 together to see what the character is *really* doing at that part."	Examples: • "Jasmine did an excellent job re-reading with an eye for evidence to support her claim." • "Let's work to connect that evidence with the claim."

A2. Rubric for the "Redirecting Off-task Behavior" Outcome in the Establishing Classroom Norms Teaching Scenario

Low 1, 2, 3	Mid 4, 5, 6	High 7, 8, 9, 10
The teacher struggled to facilitate a discussion about classroom norms. At the low end the discussion was derailed because the teacher did not effectively redirect student behavior. At the high end, the discussion proceeded haltingly. **Evidence that supports a low score includes:**	**The teacher facilitated a discussion about classroom norms and their rationale, though they somewhat struggled to do so.** **Evidence that supports a mid-range score includes:**	**The teacher facilitated a well-paced discussion focused on classroom norms and their rationale.** **Evidence that supports a high score includes:**

● The teacher did not effectively redirect student behaviors. At the low end, the teacher ignores most off-task behavior. At the high end, the teacher may acknowledge a few of the off-task behaviors. However, their attempts to address off-task behavior are ineffective.
● Disruptions were lengthy and detracted from the overall quality of the discussion.
● Most dialogue/sounds are (a) teacher candidate engaging with students about their behavior (b) teacher candidate attempts to keep the discussion afloat during unaddressed student off-task behavior (c) student off-task behavior.
● The teacher did not support student engagement. Therefore, there were few moments where <u>all</u> students were engaged in on-task behavior. Some students may be engaging in off-task behavior. Others may be compliant, but not actively engaged.
● Overall teacher affect is not calm and may grow less so over the course of the simulation. Teacher may appear nervous, flustered, or rattled. Alternatively, teachers may become visibly irritated or angry
	● The teacher attempted, with mixed success, to redirect off-task student behavior.
	● Some disruptions detracted from the overall quality of the discussion.
	● Dialogue is sometimes focused on classroom norms and their rationale. At other times it is
	  ○ (a) focused on students and their misbehavior,
	  ○ (b) used in attempts to keep the discussion afloat during unaddressed student misbehavior (ex: repeating the question several times or to different students) or
	  ○ (c) off topic (ex: teacher gets "sucked in" to student misbehavior and ends up talking at length about rocket club, plastic man, fun zone, or any of the other topics the students bring up as part of their off-task behaviors.
	● The teacher inconsistently supported student engagement. <u>All</u> students were on-task and demonstrating signs of engagement for some of the simulation.
	● Teacher affect is, for the most part, calm though there may be occasional instances where they appear otherwise. The length and intensity of these "non-calm" moments should move the score up or down.
		● The teacher efficiently redirected all off-task student behavior.
		  ○ At the high end, the teacher uses positive behavioral praise.
		● Disruptions were extremely brief and did not detract from the overall quality of the discussion.
		  ○ The faster and less invasive (ex: a quick non-verbal) the redirections, the higher the score.
		● Most of the dialogue advances the classroom discussion about norms.
		● The teacher supported student engagement such that <u>all</u> students were engaged in on-task behavior for the majority (low end) to almost all (high end) of the simulation.
		● Teacher affect is calm, and they remain unruffled throughout the simulation.

A3. Rubric for the "Supporting Text-Focused Instruction" Outcome

Low 1, 2, 3	Mid 4, 5, 6	High 7, 8, 9, 10
At the low end, teacher responses to student contributions (or lack thereof) may have led to confusion or incorrect interpretations of the text. At the high end, teacher responses to student contributions did not actively support understanding of this text. That is, we can infer that student comprehension was likely unaffected by teacher responses to student contributions. Evidence that supports a low score includes: ● Teacher responses to student contributions, when provided, are largely perfunctory in nature. Though, at the high end of low-range scores, teachers may occasionally probe, re-voice, or extend student responses. ● Teacher treats all responses as equally valid. Therefore, when students provide responses that are not supported by the text, the teacher either ignores or affirms them.	**We can infer that teacher responses to student contributions supported student understanding of this text. We can infer that because of teacher responses to student contributions, student comprehension was greater than it would have been were students reading this passage on their own.** Evidence that supports a mid-range score includes: ● Most teacher responses to student contributions are non-perfunctory although there are occasional examples of perfunctory feedback. ● Teacher consistently uses probing (text-based and non-text-based) to prompt students to justify their responses (e.g., "What makes you think that?" "Can you point us to a section in the text that gave you that idea"). ● While the teacher consistently probes students for textual evidence, the teacher does not push students to provide a text-based warrant (the link between their text-based evidence and their claim). At the high end of mid-range scores, the teacher may probe students for a warrant, but does not engage with students' contributions long enough to get them to articulate their warrant clearly. ● When a student gives a response that is not supported by the text the teacher attempts to scaffold their comprehension of the text but is not always effective at supporting struggling students in revising their responses based on the text. ● Teacher regularly restates student responses in academic language or extends student responses. ● Teacher attempts to use descriptive feedback although it may be too vague to support future student comprehension and discussion. They may label or affirm an academic behavior but do not explain <u>*why*</u> it is beneficial (e.g., "Good text-to-self connection" "I like that you thought about how the text made you feel.").	**We can infer that teacher responses to student contributions supported a nuanced understanding of this text. In addition, we can infer teacher responses to student contributions supported students in developing skills that will help them better comprehend and discuss future texts.** Evidence that supports a high score includes: ● Almost all responses to student contributions are non-perfunctory. ● Teacher seeks to scaffold students' understanding of the text rather than treating all responses as equally valid (e.g. "So you say Lisa was excited about the lie detector test, but I see several instances that might contradict that claim. Let's reread the section. If we get to something that supports OR undermines your claim, I want you to raise your hand so we can pause and discuss."). If a student has a misconception, the teacher supports them with multiple feedback loops, until they revise their response. ● Teacher sustains interactions with students that we can infer deepen their understanding about the text through feedback loops (e.g. After probing a student for textual evidence, "I like that you provided evidence from the text, but can you tell me how you knew that 'her heart beating fast' was related to her feeling of nervousness?" In this example, the teacher stays with this idea, asking multiple follow-up questions, rather than moving on after a single probe.) ● Multiple instances of high-quality descriptive feedback where the teacher explicitly highlights a specific part of a student's response <u>*and*</u> explains its value (e.g., "Ava just made a text-to-self connection when she described the way her heart pounded the time she lied to her sister. If we had only read the words 'heart pounding' it might be hard to know if Lisa was nervous or excited. However, when we think about when we have been in situations similar to characters, like Ava did, we can use our memory of the physical and emotional things we experienced to better make sense of the information the text provides about a character."). ● Teacher supports students in arriving at a *complete* answer to each discussion question. That is, answers include a claim, textual evidence, and a warrant. At the low end of the high profile, a teacher may only successfully support students in doing this for one student response. At the high end, the teacher supports students in providing evidence and a warrant for every viable claim offered by students. ● At the high end, teacher synthesizes student contributions (e.g "So Dev suggested Lisa was

		a spy, while Savannah thought she might be a reporter. Because we could find examples of Lisa acting in a way that is consistent with the behavior and motivation of spies *and* journalists, we realized that both responses are justified using evidence from the text.") and/or coordinates student responses with each other ("When Ethan said X, it echoed Jasmine's idea about Lisa…").

A4. Descriptive statistics for baseline and pretest measures (including reliability alphas)

		Teacher candidate sample (Studies 1, 2, 3, & 5)		Undergraduate participant sample (Study 4)	
	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
Pretest Quality of Pedagogical Performance	2-10	5.62	0.75-0.88	4.36	0.86

Table A5. Demographic characteristics across studies and joint tests of statistical significance for means across studies.

		Effect Size Differences			
	Bench mark	Multiple Cohort (Time)	Switching Replication (Task)	Conceptual Replication (Population)	Conceptual Replication (Delivery)
	Study 3	*vs* Study 1	*vs* Study 2	*vs* Study 4	*vs* Study 5
	(1)	(2)	(3)	(4)	(5)
GPA	**3.43 [0.34]**	0.01	0.05	0.07	0.34
% Either parent a teacher	**0.35 [0.46]**	-0.44*	0.02	-0.07	-0.23
% Mother education-college or above	**0.96 [0.34]**	-0.53*	0.03	-0.10	-0.77*
% Father education-college or above	**0.83 [0.39]**	-0.50*	0.06	0.07	-0.63
% Female	**0.69 [0.39]**	0.33	0.03	-0.64*	-0.03
% Over the age of 21	**0.86 [0.50]**	0.24	-0.02	-0.23*	-0.14*
% White	**0.83 [0.44]**	-0.34	0.05	-0.31+	-0.42
Location of high school attended					
% Rural	**0.26 [0.39]**	-0.08	-0.32	-0.17	-1.78**
% Suburban	**0.75 [0.48]**	-0.22	-0.74*	0.09	-1.85**
% Urban	**0.01 [0.29]**	0.11	0.13	0.23	0.72**
Average SES of high school attended					
% Low SES	**0.05 [0.22]**	0.08	0.04	0.04	-0.19
% Middle SES	**0.77 [0.48]**	-0.42	0.01	-0.07	-1.34**
% High SES	**0.23 [0.41]**	0.02	-0.02	0.24	0.39**

Majority race of high school attended					
% Primarily students of color	**0.07 [0.20]**	0.00	0.04	0.16	-0.32
% Mixed	**0.46 [0.49]**	-0.09	0.05	-0.03	-2.67**
% Primarily white students	**0.56 [0.50]**	-0.26	-0.03	0.05	-0.64*
Average achievement level of high school attended					
% Primarily low achieving	**0.04 [0.14]**	0.17	0.00	0.11	0.11
% Primarily middle achieving	**0.57 [0.50]**	-0.38	0.07	-0.14	-0.56
% Primarily high achieving	**0.39 [0.49]**	0.05	-0.03	0.26	0.09
Instructional quality performance score at pretest	**3.43 [1.33]**	0.14	0.37*	-0.58 **	-0.74**
F-statistic		38.88	180.62	97.86	596.63
P-value for *F*-test		0.00	0.00	0.00	0.00

Notes: Regression-adjusted means and standard deviations in brackets are presented in Column 1 for Study 3. Columns (2)- (5) present standardized effect size differences for each covariate between Study 3 and the relevant study. 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. Results of the joint f-test are presented in the last row for each set of tests. +p < .10. *p < .05. **p < .01.

Table A6. Conceptual Replication Design, Causal Replication Assumptions, and Results of Diagnostics for Addressing Assumptions

		Replication Design Assumptions (R1-R2)		**Individual Study Assumptions (S1-S3)**		
Source of Variation	**Replication Design**	*R1. Treatment & Outcome Stability Across Studies*	*R2. Equivalent Causal Estimand Across Studies*	*S1. Unbiased Identification of Treatment Effects*	*S2. Unbiased Estimation of Treatment Effects*	*S3. Correct and Accurate Reporting of Treatment Effects*
Timing of Study	Multiple Cohort (Study 3 vs. Study 1)	Treatments ✓ Outcomes ✓	Participants (×) Settings ✓ Causal quantity ✓ Time ×	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Teaching Task	Switching Replication (Study 3 vs. Study 2)	Treatments ✓ Outcomes ×	Participants (×) Settings × Causal quantity ✓ Time (×)	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Mode of Delivery	Conceptual Replication with online vs. in-person (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 ✓
Population and Setting Characteristics	Conceptual Replication with Different	Treatments ✓ Outcomes ✓	Participants × Settings × Causal quantity ✓ Time (×)	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓

	Units and Settings (Study 3 vs. Study 4)					

Table A6 summarizes: 1) the replication design (R1-R2) and individual study (S1-S3) for direct replication of results across studies (Author, 2019; Author, 2021); 2) the source of the variation under investigation; 3) the replication design for evaluating the source of the variation; and 4) results from the diagnostic assessments for each assumption. ✓ indicate that we judge the assumption has been met by either looking at diagnostic results (e.g. balance tests for an RCT), conducting sensitivity checks (e.g. robustness of results over multiple model specifications, and data analysts) or by study design (e.g. randomization). A bolded × indicates that the study feature or assumption was systematically varied; (×) indicates variation across studies that was unplanned or required for feasibility reasons. These variations represent alternative explanations for why study results may not replicate. To address potential differences in the composition of participant characteristics across studies, we also assess the replicability of results with an adjusted sample that has been reweighted to appear observationally similar to the benchmark Study 3 sample.

A7. Balance table for full and analytic samples for Study 1 (Spring 2018)

	Study 1 (Spring 2018)			
	Full sample		Analytic sample	
	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)
	(1)	(2)	(3)	(4)
Baseline demographics				
GPA	3.46	-0.05	3.46	-0.07
% Either parent a teacher	0.26	-0.07	0.26	-0.06
% Mother education- college or above	0.91	-0.12+	0.91	-0.12+
% Father education- college or above	0.76	-0.11	0.76	-0.12
% Female	0.77	-0.06	0.78	-0.07
% Over the age of 21	0.82	0.20*	0.83	0.19*
% White	0.75	0.01	0.75	0.02
Location of high school attended				
% Rural	0.23	-0.01	0.22	-0.01
% Suburban	0.78	-0.01	0.79	0.01
% Urban	0.00	0.03	0.00	0.01
Average SES of high school attended				
% Low SES	0.06	0.00	0.06	-0.01
% Middle SES	0.65	-0.03	0.64	-0.03
% High SES	0.30	0.05	0.31	0.05
Majority race of high school attended				
% Primarily students of color	0.09	-0.03	0.09	-0.02
% Mixed	0.49	-0.09	0.51	-0.09
% Primarily white students	0.46	0.13	0.44	0.13
Average achievement level of high school attended				
% Primarily low achieving	0.08	0.00	0.07	0.00
% Primarily middle achieving	0.40	0.09	0.41	0.09
% Primarily high achieving	0.52	-0.09	0.52	-0.09
Instructional quality performance score at pretest	3.67	3.56	3.71	3.60
Attrition rate (from initial sample)	0%	+4%	3%	+4%

Notes: Demographic information comes from data collected by the teacher preparation program for Study 1. Each row represents regression-adjusted means from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks. The attrition rate in columns 1 and 3 represent the attrition rates from the initial randomization samples in the control group for the full and analytic samples; the attrition rate in columns 2 and 4 represent the difference in attrition rates between the control and treatment groups for the full and analytic samples. +p < .10. *p < .05. **p < .01.

Table A8. Balance table for full and analytic samples for Study 2 (Fall 2018)

	Study 2 (Fall 2018)			
	Full sample		Analytic sample	
	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)
Baseline demographics				
GPA	3.49	-0.03	3.48	-0.03
% Either parent a teacher	0.31	0.07	0.30	0.07
% Mother education- college or above	0.96	0.00	0.96	0.00
% Father education- college or above	0.81	0.04	0.81	0.04
% Female	0.70	0.00	0.70	0.00
% Over the age of 21	0.89	-0.07	0.89	-0.07
% White	0.83	-0.02	0.83	-0.02
Location of high school attended				
% Rural	0.23	0.02	0.22	0.02
% Suburban	0.69	-0.02	0.71	-0.02
% Urban	0.09	-0.01	0.09	-0.01
Average SES of high school attended				
% Low SES	0.07	-0.03	0.07	-0.03
% Middle SES	0.83	-0.02	0.82	-0.02
% High SES	0.17	0.06	0.18	0.06
Majority race of high school attended				
% Primarily students of color	0.07	0.02	0.07	0.02
% Mixed	0.50	0.00	0.51	0.00
% Primarily white students	0.54	0.00	0.53	0.00
Average achievement level of high school attended				

% Primarily low achieving	0.04	0.00	0.03	0.00
% Primarily middle achieving	0.66	-0.10	0.67	-0.10
% Primarily high achieving	0.29	0.10	0.29	0.10
Instructional quality performance score at pretest	3.98	3.95	3.94	3.94
Attrition rate (from initial sample)	3%	+7%	14%	-1%

Notes: Demographic information comes from data collected by the teacher preparation program for Study 2. Each row represents regression-adjusted means from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks. The attrition rate in columns 1 and 3 represent the attrition rates from the initial randomization samples in the control group for the full and analytic samples; the attrition rate in columns 2 and 4 represent the difference in attrition rates between the control and treatment groups for the "full" and "analytic" samples. +p < .10. *p < .05. **p < .01

Table A9. Balance table for full and analytic samples for Study 3 (Spring 2019)

	Study 3 (Spring 2019)			
	Full sample		Analytic sample	
	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)
Baseline demographics				
GPA	3.41	0.03	3.40	0.02
% Either parent a teacher	0.37	-0.08	0.34	-0.08
% Mother education- college or above	0.97	-0.03	0.96	-0.03
% Father education- college or above	0.81	0.05	0.84	0.01
% Female	0.65	0.08	0.63	0.09
% Over the age of 21	0.85	0.07	0.86	0.06
% White	0.86	-0.06	0.86	-0.08
Location of high school attended				
% Rural	0.26	0.01	0.24	0.01
% Suburban	0.80	-0.01	0.80	0.01
% Urban	-0.03	-0.01	-0.02	-0.02
Average SES of high school attended				
% Low SES	0.02	0.06	0.03	0.07
% Middle SES	0.90	-0.19*	0.86	-0.17+

% High SES	0.16	0.13	0.19	0.10
Majority race of high school attended				
% Primarily students of color	0.06	0.00	0.07	0.00
% Mixed	0.50	-0.05	0.54	-0.05
% Primarily white students	0.55	0.04	0.50	0.05
Average achievement level of high school attended				
% Primarily low achieving	0.06	-0.04	0.04	-0.02
% Primarily middle achieving	0.52	0.11	0.54	0.10
% Primarily high achieving	0.42	-0.08	0.42	-0.08
Instructional quality performance score at pretest	3.77	3.33	3.60	3.30
Attrition rate (from initial sample)	5%	0%	15%	+2%

Notes: Demographic information comes from data collected by the teacher preparation program for Study 3. Each row represents regression-adjusted means from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks. The attrition rate in columns 1 and 3 represent the attrition rates from the initial randomization samples in the control group for the full and analytic samples; the attrition rate in columns 2 and 4 represent the difference in attrition rates between the control and treatment groups for the "full" and "analytic" samples. +p < .10. *p < .05. **p < .01.

Table A10. Balance table for full and analytic samples for Study 4 (Fall 2019)

	Study 4 (Fall 2019)			
	Full sample		Analytic sample	
	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)
Baseline demographics				
GPA	3.40	0.17	3.42	0.18
% Either parent a teacher	0.34	-0.02	0.35	-0.02
% Mother education- college or above	0.81	0.11	0.82	0.08
% Father education- college or above	0.80	0.12+	0.81	0.10
% Female	0.52	0.11	0.53	0.10
% Over the age of 21	0.58	-0.05	0.60	-0.02
% White	0.61	0.00	0.62	-0.01

Location of high school attended				
% Rural	0.06	0.01	0.07	0.01
% Suburban	0.84	0.00	0.83	0.00
% Urban	0.10	-0.01	0.11	-0.01
Average SES of high school attended				
% Low SES	0.09	-0.04	0.09	-0.04
% Middle SES	0.60	0.02	0.60	0.03
% High SES	0.31	0.02	0.31	0.01
Majority race of high school attended				
% Primarily students of color	0.12	-0.04	0.11	-0.02
% Mixed	0.34	0.02	0.36	0.03
% Primarily white students	0.54	0.01	0.52	-0.01
Average achievement level of high school attended				
% Primarily low achieving	0.07	0.00	0.07	0.00
% Primarily middle achieving	0.33	0.11	0.32	0.12
% Primarily high achieving	0.61	-0.11	0.61	-0.12
Instructional quality performance score at pretest	3.00	2.72	3.02	2.73
Attrition rate (from initial sample)	14%	0%	18%	-1%

Notes: Demographic information comes from data collected by the research team for Study 4. Each row represents regression-adjusted means from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks. The attrition rate in columns 1 and 3 represent the attrition rates from the initial randomization samples in the control group for the full and analytic samples; the attrition rate in columns 2 and 4 represent the difference in attrition rates between the control and treatment groups for the "full" and "analytic" samples. +p < .10. *p < .05. **p < .01.

Table A11. Balance table for full and analytic samples for Study 5 (Spring 2020)

	Study 5 (Spring 2020)			
	Full sample		Analytic sample	
	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)	Control Group (Self-reflection) Mean (SE)	Coaching/ Control group difference Coefficient (SE)
Baseline demographics				
GPA	3.49	0.04	3.49	0.04
% Either parent a teacher	0.35	-0.03	0.34	-0.03

% Mother education- college or above	0.81	0.03	0.81	0.03
% Father education- college or above	0.71	0.10+	0.71	0.11+
% Female	0.70	-0.06	0.70	-0.02
% Over the age of 21	0.86	-0.14+	0.88	-0.16*
% White	0.76	-0.09	0.75	-0.10
Location of high school attended				
% Rural	0.12	0.02	0.10	0.02
% Suburban	0.39	-0.07	0.41	-0.08
% Urban	0.43	0.07	0.43	0.09
Average SES of high school attended				
% Low SES	0.02	0.02	0.02	0.02
% Middle SES	0.25	-0.02	0.24	-0.03
% High SES	0.54	-0.05	0.55	-0.05
Majority race of high school attended				
% Primarily students of color	0.04	0.02	0.04	0.02
% Mixed	0.20	-0.04	0.21	-0.05
% Primarily white students	0.44	-0.06	0.43	-0.07
Average achievement level of high school attended				
% Primarily low achieving	0.04	0.02	0.03	0.02
% Primarily middle achieving	0.48	-0.03	0.49	-0.03
% Primarily high achieving	0.48	0.01	0.48	0.01
Instructional quality performance score at pretest	2.77	2.91	2.75	2.90
Attrition rate (from initial sample)	2%	-2%	4%	+8%

Notes: Demographic information comes from data collected by the teacher preparation program for Study 5. Each row represents regression-adjusted means from a separate regression with the same right-hand specification but different covariate as the dependent variable. Models include controls for randomization blocks. The attrition rate in columns 1 and 3 represent the attrition rates from the initial randomization samples in the control group for the full and analytic samples; the attrition rate in columns 2 and 4 represent the difference in attrition rates between the control and treatment groups for the "full" and "analytic" samples. +p < .10. *p < .05. **p < .01.

A12. Propensity Score Weighting for Adjusted Coaching Effects

Because of differences in study participant characteristics, we employed propensity score matching and weighting techniques to present adjusted estimates of the impact of coaching

for each study. Our process for "matching" participants across the two sets of studies involved reweighting the samples in Studies 1, 2, 4 & 5 to look more similar on observable characteristics to our benchmark sample in Study 3. In this context, propensity score matching will, for example, make the entire Study 2 sample more like the target Study 3 sample based on observable characteristics like demographics and pre-treatment survey responses. We estimate propensity scores using generalized boosted models (GBM), a data-driven, non-parametric, and replicable approach to propensity score estimation (McCaffrey et al., 2013). We estimate the GBM models using pre-treatment covariates, including whether participants were female, White, High-School GPA, 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 simulator performance, self-efficacy beliefs, multicultural attitudes towards teaching, and beliefs regarding culturally-responsive teaching. Tables A12.1-A12.4 show that while propensity score weighting resulted in balance for most covariates (with standardized mean differences < .20), key imbalances remain on beliefs regarding culturally responsive teaching and baseline simulator performance in Studies 2, 4, 5, and 5, and additionally, self-efficacy beliefs in Study 1.

Results of Propensity Weighting

Table A12.1 displays standardized mean differences between the sample in Study 3 and the sample in Study 1 prior to and after propensity score weighting, using the teacher candidate sample as the reference group. Using propensity score weighting, standardized differences in Self-efficacy Beliefs (overall), Self-efficacy Beliefs (classroom management) , and Baseline Simulator Performance are above a cutoff of .20 , which is acceptable by convention (McCaffrey et al., 2013).

Table A12.1. Standardized Mean Differences between Study 3 and Study 1

	Before Weighting	After Weighting
	Standardized difference	
% Female	-0.16	-0.14
% White	-0.17	-0.04
Location of high school attended		
% Urban	-0.11	-0.03
Average achievement level of high school attended		
% Primarily high achieving	-0.18	-0.07

Average SES of high school attended % Middle SES	0.02	0.01
Self-efficacy Beliefs (overall)	-0.32	-0.35
Teacher Multicultural Attitudes score	0.04	0.05
Culturally Responsive Teaching Self-efficacy score	-1.43	0.15
Self-efficacy Beliefs (classroom management)	-0.31	-0.31
High School GPA	-0.05	-0.07
Baseline Simulator Performance	0.06	0.46

Table A12.2 displays standardized mean differences between the sample in Study 3 and the sample in Study 2 prior to and after propensity score weighting, using the teacher candidate sample as the reference group. Using propensity score weighting, the standardized difference in just Culturally Responsive Teaching Self-efficacy score is above a cutoff of .20 , which is acceptable by convention (McCaffrey et al., 2013).

Table A12.2. Standardized Mean Differences between Study 3 and Study 2

	Before Weighting	After Weighting
	Standardized difference	
% Female	0.00	0.00
% White	0.00	0.00
Location of high school attended % Urban	0.02	0.02
Average achievement level of high school attended % Primarily high achieving	0.03	0.03
Average SES of high school attended % Middle SES	0.01	0.01
Self-efficacy Beliefs (overall)	-0.15	-0.16
Teacher Multicultural Attitudes score	-0.03	-0.03
Culturally Responsive Teaching Self-efficacy score	0.38	0.45
Self-efficacy Beliefs (classroom management)	-0.15	-0.15
High School Gpa	-0.04	-0.03
Baseline Simulator Performance	0.13	0.14

Table A12.3 displays standardized mean differences between the sample in Study 3 and the sample in Study 4 prior to and after propensity score weighting, using the teacher candidate sample as the reference group. Using propensity score weighting, standardized differences in just Culturally Responsive Teaching Self-efficacy score and Baseline Simulator Performance are above a cutoff of .20, which is acceptable by convention (McCaffrey et al., 2013).

Table A12.3. Standardized Differences between Study 3 and Study 4

	Before Weighting	After Weighting
	Standardized difference	
% Female	-0.32	-0.15
% White	-0.16	-0.20
Location of high school attended % Urban	0.04	-0.01
Average achievement level of high school attended % Primarily high achieving	-0.07	-0.19
Average SES of high school attended % Middle SES	-0.01	-0.01
Self-efficacy Beliefs (overall)	-0.06	0.00
Teacher Multicultural Attitudes score	-0.01	0.00
Culturally Responsive Teaching Self-efficacy score	0.26	2.26
Self-efficacy Beliefs (classroom management)	-0.03	0.04
High School GPA	0.07	0.06
Baseline Simulator Performance	-0.01	0.35

Table A12.4 displays standardized mean differences between the sample in Study 3 and the sample in Study 5 prior to and after propensity score weighting, using the teacher candidate sample as the reference group. Using propensity score weighting, standardized differences in just Culturally Responsive Teaching Self-efficacy score and Baseline Simulator Performance are above a cutoff of .20 , which is acceptable by convention (McCaffrey et al., 2013).

Table A12.4. Standardized Differences between Study 3 and Study 5

	Before Weighting	After Weighting
	Standardized difference	
% Female	0.00	-0.05
% White	-0.07	0.02
Location of high school attended % Urban	0.06	0.03
Average achievement level of high school attended % Primarily high achieving	-0.06	0.00
Average SES of high school attended % Middle SES	0.03	0.03
Self-efficacy Beliefs (overall)	-0.09	-0.13
Teacher Multicultural Attitudes score	0.07	0.00
Culturally Responsive Teaching Self-efficacy score	-0.02	-1.09
Self-efficacy Beliefs (classroom management)	-0.10	-0.12
High School Gpa	0.03	0.00
Baseline Simulator Performance	0.21	0.65

Appendix Table A13.
Meta-Analytic average treatment effect size and average treatment effect Size by study for modified IOWA Connors rating scale for the "establishing classroom norms" scenario.

Candidates completed a short post-simulation survey after establishing norms simulations (studies 1, 3, 4 & 5) using a modified IOWA Connor's rating scale (Waschbusch & Willoughby, 2008). The IOWA Conners Rating Scale is a widely used brief measure of inattentive-impulsive-overactive (IO) and oppositional-defiant (OD) behavior in children. The measure was modified to include minor off-task behaviors that the student avatar displayed in the simulation.

Across the four studies, the meta-analytic coaching effect was negative and statistically significant (-.49 SD, *p*-value < 0.01). The test of homogeneity did not indicate significant differences in effect estimates across studies (*Q*-statistic = 5.14; *df* = 3; *p*-value = 0.16). Columns 2-6 in Appendix Table 13 provide separate effect estimates in standard deviation units for each study. Results indicate that teacher candidates who received coaching perceived student avatars as displaying IO and OD behaviors less than those who self-reflected between. It is important to note that the student avatar behaviors were identical across groups, an important affordance of the simulation platform. We observe these significant coaching effects in Studies 1, 3, and 5, all of which included teacher candidates. However, this shift in ratings of student behavior was not replicated for Study 4 with undergraduates who were not enrolled in a teacher preparatory program. This suggests that candidates who received coaching evaluate students' behaviors as being less extreme or problematic, which we see as suggestive they are more confident in their skills as managing such behaviors. Just as teacher candidates benefit more from coaching than undergraduates in terms of their observable skill development, they also seem more likely to change their views about children's behaviors than their coached undergraduate counterparts.

	Meta-analytic Treatment effect (1)	Study 1 Treatment effect (2)	Study 3 Treatment effect (3)	Study 4 Treatment effect (4)	Study 5 Treatment effect (5)
Modified IOWA Connors Scale	-0.49*	-0.69**	-0.66**	-0.24	-0.31*
	(0.09)	(0.16)	(0.18)	(0.20)	(0.17)
Q-statistic	5.14				
Analytic sample N		99	94	90	104

Notes: Adjusted coaching effects are reported in each column. Coefficients and standard errors (in parentheses) are reported 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 (1) represents the overall meta-analytic coaching effect across the five studies. Models for each study-specific effect include controls for randomization blocks, participants' gender, race, high school GPA, baseline score, indicators for missing baseline scores, and interactor fixed-effects. +p < .10. *p < .05. **p < .01