ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-468

Experimental Evidence on the Robustness of Coaching Supports in Teacher Education

Julie Cohen
University of Virginia

Anandita Krishnamachari
University of Virginia

Vivian C. Wong
University of Virginia

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.

VERSION: September 2021

Suggested citation: Cohen, Julie, Anandita Krishnamachari, and Vivian C. Wong. (2021). Experimental Evidence on the Robustness of Coaching Supports in Teacher Education. (EdWorkingPaper: 21-468). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/dgf9-ca95

Experimental Evidence on the Robustness of Coaching Supports in Teacher Education

Julie Cohen
Anandita Krishnamachari
Vivian C. Wong
University of Virginia

September 2021

Authors are listed in alphabetical order to denote equal contribution by each co-author.

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.

Experimental Evidence on the Robustness of Coaching Supports in Teacher Education

There is considerable evidence that teachers improve dramatically in their early years of classroom experience (Atteberry, Loeb & Wyckoff, 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, Hammerness & McDonald, 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 (Cohen et al., 2020).

Coaching is a promising option for expediting skill development during approximations (Kraft, Blazar, & Hogan, 2018). Theory suggests that coaches serve as experts who can observe teachers, evaluate strengths and weaknesses, and develop individualized strategies to promote improvement (Kraft et al., 2018; Coburn & Woulfin, 2012). Coaching is increasingly common for in-service teachers (Stahl, Sharplin & Kehrwald, 2016) and has been shown to improve teachers' attitudes toward teaching, feelings of self-efficacy, instructional skills, and student achievement (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. 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). 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., Bowman & McCormick, 2000).

In earlier work, we found compelling experimental evidence that directive coaching could dramatically improve candidates' classroom management skills (Cohen, Wong

Krishnamachari, & Berlin, 2020). What is less clear is whether such findings would replicate in additional experimental evaluations with systematic variations in participant characteristics, settings, and times. Answering such questions can help teacher educators understand the contexts and conditions under which coaching is most effective during the all too brief timeframe of pre-service preparation (Blazar & Kraft, 2015; Hill, Beisiegel, & Jacob, 2013; Ronfeldt, 2015).

Through a series of conceptual replication studies, this paper examines the robustness of effects from a standardized coaching protocol for improving teaching in mixed reality simulation (MRS) settings. The simulation platform features a virtual classroom and student avatars who are remotely controlled by an actor trained to facilitate realistic classroom interactions. The replications were designed to introduce sources of variation across studies to evaluate the robustness of coaching effects across different time periods, teaching tasks, participant characteristics and course experiences, and delivery modes of practice and coaching. Given the limited duration but critical nature of teacher preparation, the field needs more rigorous evidence about the best ways to support candidates.

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; Grossman, Ronfeldt, & Cohen, 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 their classroom observations (Grossman et al., 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 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 that 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, Rodriguez, Lignugaris/Kraft, Hynes & Hughes, 2014).

Simulations are also useful for conducting causal work 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 the development of specific instructional skills while limiting other sources of variation, such as the content of instruction (Cohen, Ruzek, & Sandilos, 2018; Cohen, 2018), the influence of mentors (Goldhaber, Krieg, Naito & Theobald, 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, practice sessions in simulated classrooms alone are likely not enough to help candidates develop and improve their 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 (Deussen, Coskie, Robinson & Autio, 2007; Hammond & Moore, 2018). This type of directive coaching can also help candidates understand the impact of instruction on students (Cohen et al., 2020).

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 practice in the simulation and diagnoses the candidate's 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 teaching 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 sds), as well as on their perceptions of the student avatars (Cohen et al., 2020).

Robustness of Coaching Effects across Key Sources of Variation

While these experimental results suggest that directive coaching *can* improve candidates' pedagogical practice, 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 conducted a series of replication studies to examine the replicability of effects across four key sources of systematic effect variation:

(1) Timing of study. We want a coaching model that "works" across multiple years of administration, 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 efficacious 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. In this study, we assess the replicability of coaching effects over multiple cohorts of candidates at the same institution.

(2) Teaching task. Most teachers are not equally skilled across teaching domains (Cohen, Raudenbush, & Ball, 2003). A teacher might be strong at establishing warm relationships with students but struggle with providing clear and accurate instructional explanations (Cohen, 2018; Pianta & Hamre, 2009; Hill, Ball & Schilling., 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 this replication effort, we compare the impacts of coaching on two types of pedagogical skills: managing students' off-task behaviors while "establishing norms" and "providing feedback" during text-based discussions.

(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. Given that online coaching programs could provide a more resource-effective way of supporting larger numbers of candidates (Israel, Knowlton, Griswold & Rowland, 2009; Rock et al., 2013; Stapleton, Tschida & Cuthrell, 2017), we examine the replicability of coaching effects over two different modalities of delivery – online Zoom sessions vs. in-person.

(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 they note that it is difficult to disentangle the two because most coaching programs are accompanied by additional training. In our context, candidates practice and receive coaching feedback on pedagogical skills that are discussed in their 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.

Research Methods

To examine the robustness of coaching effects in MRS settings, we use data from five randomized control trials (RCTs) that introduced systematic variations across studies on the dimensions noted above. Figure 1a describes the timing of each of the five individual replication 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.

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 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 individual study, participants were randomly assigned within course sections to receive coaching or 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. 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 A5-A9). 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, Imbens, & Rubin, 1996). For each of the studies, attrition was minimal (less than 15%), with no evidence of differential attrition between groups (see Appendices A5-A9 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 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 focused on "establishing classroom norms" while redirecting off-task behaviors;

in Study 2, participants focused on "providing high-quality feedback" while leading a text-based 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.

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 Appendix A4 for how implementation fidelity of the coaching protocol was assessed).

Self-reflection (business-as-usual) condition. Instead of receiving coaching feedback after Time 2, participants in the control condition engaged in a researcher-designed "self-reflection protocol" that asked participants to identify perceived strengths and weakness, 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 A2 for a descriptive summary of baseline measures of participant characteristics and their psychometric properties and outcomes).

Pedagogical outcomes. Our primary outcome measures were obtained from observational protocols designed by the research team to assess the quality of participants' instructional skills in simulation sessions. Rubrics for the "setting classroom norms" scenario were based on the Responsive Classroom (2014) framework, while rubrics for the "feedback during a text-based discussion" were derived from relevant literature about high-quality feedback practices (e.g., Hattie & Timperley, 2007). 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. In this paper, the primary outcome of interest in the replication studies is a measure of "overall quality of pedagogical performance" (Cohen et al., 2020). Scores for the instructional performance measure ranged from 1-low to 10-high, and reflected the extent to which the goals of teaching task were met (i.e. leading a text-based discussion or establishing classroom norms).

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 (Steiner, Wong, & Anglin, 2020; Wong, Anglin, & Steiner, 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, Cook & Campbell 2002).[1] We compared coaching effects for two intervention intervals, where candidates practiced "leading a text-based discussion" (Study 2) in the first period and "managing off-task student behaviors" (Study 3) in the second. 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 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.

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

[1] 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.

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. 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 3).

Next, we estimated the *overall average treatment effect* across the five studies using a fixed effects meta-analytic approach, where each study's effect size was weighted by the inverse variance of the effect estimate (see Table 3). Finally, to evaluate *effect heterogeneity* across studies, we examined the *Q*-statistic for the test of homogeneity (Hedges & Schauer, 2018). If the Null hypothesis is rejected and effect heterogeneity is inferred, we compared 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). We assessed replication success by comparing the direction, magnitude, and statistical significance patterns of effects, as well as by conducting statistical tests of difference and equivalence in effect size estimates (Steiner & Wong, 2018).

Results

Diagnostic Results of Replication Assumptions

Table 1 summarizes participant and setting characteristics for each study to demonstrate the extent to which these systematically varied and/or replicated across studies. For Studies 1, 2,

3, and 5, teacher candidate samples were similar in demographic characteristics, and the coaching sessions were delivered with consistent adherence to the standardized protocol (see Appendix A4). The bolded text highlights systematic differences in participant and setting characteristics introduced across replication efforts. In particular, the target population and setting in Study 4 differed from the other replication studies. Participants in Study 4 had different undergraduate course experiences (and less training in pedagogical methods), were younger, more male, and less white than candidates in Studies 1, 2, 3, and 5. Appendix A3 summarizes assumptions for each replication design and the results of our diagnostic assessments of the extent to which assumptions were met.

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. Across the five studies, the meta-analytic coaching effect was positive, large, and statistically significant (1.44 SD, p-value < 0.01; Column 1, Table 2). However, the test of homogeneity indicated significant differences in effect estimates across studies (Q-statistic = 8.12; df = 4; p-value = 0.09). Columns 2-6 in Table 2 provide separate effect estimates in standard deviation units for each study. Effect sizes ranged from 0.58 SD (p-value = *ns*; Study 4) to 1.67 SD (p-value < 0.01; Study 5). Coaching effects for candidate samples (Studies 1, 2, 3, and 5) were consistently large and statistically significant (ranging from 1.34 SD in Study 2 to 1.67 SD in Study 5), while the coaching effect for undergraduates (Study 4) was 0.58 SD and not statistically significant.

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 ✓ indicating replication success by a pre-specific criterion (magnitude, sign, statistical significance pattern of results, and no statistical difference), and X indicating replication failure by the pre-specified criterion.

Timing of study. Results from Table 3 show that coaching effects were robust across variation in study timing by each criterion. The coaching effect for Study 1 was 1.65 SD (p-value < 0.01) while the coaching effect was 1.37 SD (p-value < 0.01) for Study 3. Effects were replicated in terms of direction, magnitude, and statistical significance patterns; they were not statistically different from each other ($\delta = -0.29$, p-value = 0.27).

*Teaching task***.** When participants practiced "establishing classroom norms," coaching improved their performance by 1.37 SDs (p-value <0.01, Study 3); when they practiced "providing feedback," coaching improved performance by 1.34 SDs (p-value < 0.01, Study 2). Coaching effects also were comparable in terms of magnitude, direction, and statistical significance patterns, and were not statistically different ($\delta = 0.03$, p-value = 0.94).

Mode of delivery. The coaching effect for the face-to-face sessions was 1.37 SD (p-value <0.01, Study 3) while the coaching effect for sessions delivered on Zoom was 1.67 SD (p-value<0.01, Study 5). Again, the pattern of effects was similar in terms of direction, magnitude, and statistical significance, and effects were not statistically different ($\delta = 0.29$, p-value=0.37).

Target Population and Concurrent Coursework. Finally, for teacher candidates who were enrolled in methods classes, the coaching effect was 1.37 SDs (p-value <0.01, Study 3), but for undergraduates not enrolled in preparatory courses, the effect was smaller and not statistically significant (0.58 SDs; Study 4). The effect estimates across the two studies were statistically different ($\delta = -0.79$, p<0.05).

Because samples in Studies 3 and 4 differed by *both* their participant characteristics and coursework experiences, we examined the replicability of effects after adjusting for observed participant characteristics (e.g. gender, race-ethnicity, mother's education, and type of high school attended). Even after controlling for these demographic characteristics, large and statistically significant differences in coaching effects remained (1.34 SDs for *Adjusted* Study 3 vs. 0.58 for *Adjusted* Study 4) – providing evidence that the benefit of coaching was moderated by participants' concurrent course experiences and not their demographic characteristics.

Discussion and Implications

Teacher preparation needs more evidence, particularly causal evidence, about methods 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, Pianta, Gregory, Mikami & Lun, 2011; 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 (Cohen et al., 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 experimental studies that evaluate the impact of directive coaching, using simulated classrooms to both approximate and assess teaching. Across four studies, we see significant improvements in performance 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 of longstanding 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.

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 rigorous causal and generalizable 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 (Wong et al., 2021). Our findings suggest that coaching significantly improves participants' 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. Undergraduates did not improve as much from coaching 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., 2009). 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 (Kraft et al., 2018; Scheeler, Bruno, Grubb & Seavey, 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.

An important next step in this work will be to examine whether these robust coaching effects extend to other teacher education contexts with more diverse teacher 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. 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 (Boguslav & Cohen, 2021). 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.

References

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.

Anglin KL, Wong VC, Boguslav A. A Natural Language Processing Approach to Measuring Treatment Adherence and Consistency Using Semantic Similarity. *AERA Open*. January 2021. doi:10.1177/23328584211028615

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.

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

Clemens, M. A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, *31*(1), 326-342.

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.

Cohen, J. (2018). Practices that cross disciplines?: Revisiting explicit instruction in elementary mathematics and language arts, *Teaching and Teacher Education, 69*(3), 324-335.

Cohen, J. , Ruzek, E., & Sandilos, L. (2018). Does teaching quality cross subjects?: Understanding consistency in elementary teacher practice across subjects, *AERA-Open.*

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208-231.

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

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.

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*.

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.

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.

Grossman, P., Ronfeldt, M., & Cohen, J. (2011) The power of setting: The role of field experience in learning to teach. In K.R. Harris, S. Graham, & T. Urdan (Eds.), *Educational Psychology Handbook: Vol. 4.* Washington, DC: American Psychological Association.

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.

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.

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.

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., & 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.

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.

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.

Lofthouse, R., Leat, D., Towler, C., Hallet, E., & Cummings, C. (2010). Improving coaching: evolution not revolution, research report.

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.

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. Retrieved October*, *30*, 2018.

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.

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

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.

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.

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

Wong, V.C. & Steiner, P.M. (2018). Designs of Empirical Evaluations of Non-Experimental Methods in Field Settings. *Evaluation Review*. Advance online publication. https://doi.org/10.1177/0193841X18778918.

Steiner, P.M., Wong, V.C. & Anglin, K.* (2019). A Causal Replication Framework for Designing and Assessing Replication Efforts. *Zeitschrift fur Psychologie. Vol 226, No 3.*

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

Wong VC, Anglin K, Steiner PM. Design-Based Approaches to Causal Replication Studies. Prev Sci. 2021 Jul 1. doi: 10.1007/s11121-021-01234-7. Epub ahead of print. PMID: 34212299.

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

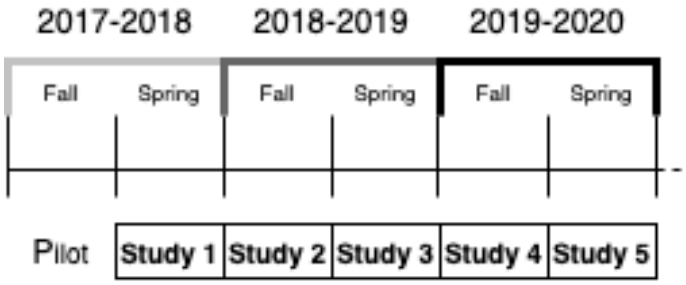
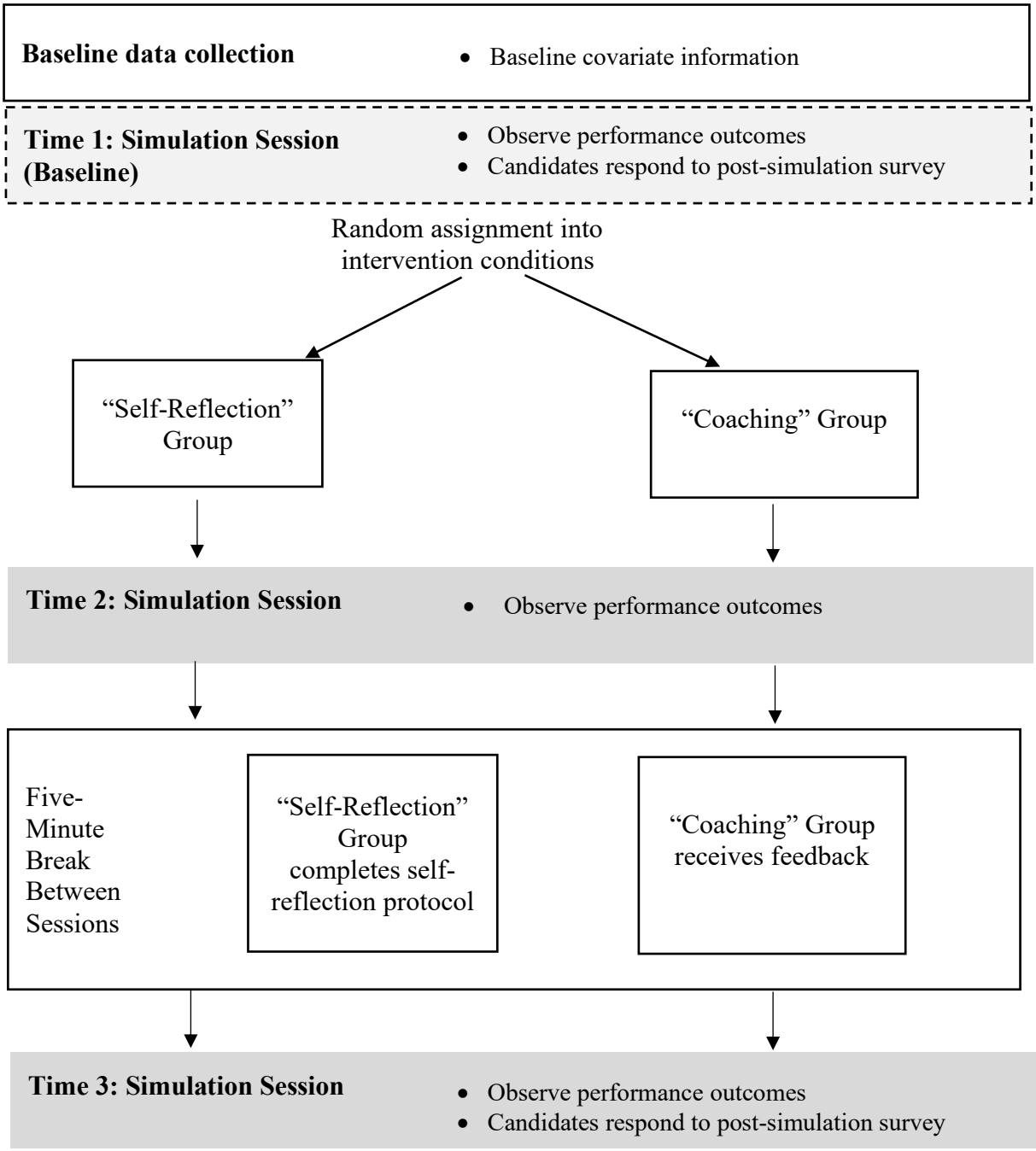


Figure 1b: Conceptual replication designs for understanding sources of systematic variation

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. Providing feedback)	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 (Wong et al., 2021; Steiner et al., 2019). See Appendix 2 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)	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	P**roviding feedback**	**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 covariate 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 Anglin et al. (2021); a higher score indicates higher similarity to a benchmark scripted treatment protocol. To examine the validity of the RCT, the research team examined baseline equivalence on an array of baseline characteristics for each study. 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: Meta-Analytic Average Treatment Effect Size, and Average Treatment Effect Size by Study

	Meta-analytic Treatment effect (1)	Study 1 Treatment effect (2)	Study 2 Treatment effect (3)	Study 3 Treatment effect (4)	Study 4 Treatment effect (5)	Study 5 Treatment effect (6)
Overall Quality of Pedagogical Performance	1.44**	1.65**	1.34**	1.37**	0.58	1.67**
	(0.11)	(0.22)	(0.24)	(0.23)	(0.36)	(0.20)
Control Mean	4.46 [1.55]	4.38 [1.60]	4.93 [1.05]	4.30 [1.88]	4.02 [1.05]	4.61 [1.32]
Q-statistic	8.12*					
Analytic sample N		99	99	95	95	104

Notes: Adjusted coaching effects are reported in each column. Coefficients and standard errors (in parentheses), and control group means and standard deviations [in brackets] are reported in columns (1) through (6) 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. One challenge with standard meta-analytic approaches for synthesizing results is that the method assumes independence in effect size estimates across studies. However, because participants in the switching replication design were shared across studies, effect estimates for Studies 2 and 3 were correlated. We addressed this dependency by first estimating the 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 to account for the correlation in effect estimates. The multivariate meta-analysis was conducted using Metafor, which allows users to specify the covariance-variance matrix for effect estimates included in the meta-analysis. Models for each study-specific effect include controls for randomization blocks, participants' gender, race, high school GPA, baseline score 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. +$p < .10$. *$p < .05$. **$p < .01$

Table 3: Replication success across series of systematic replication studies

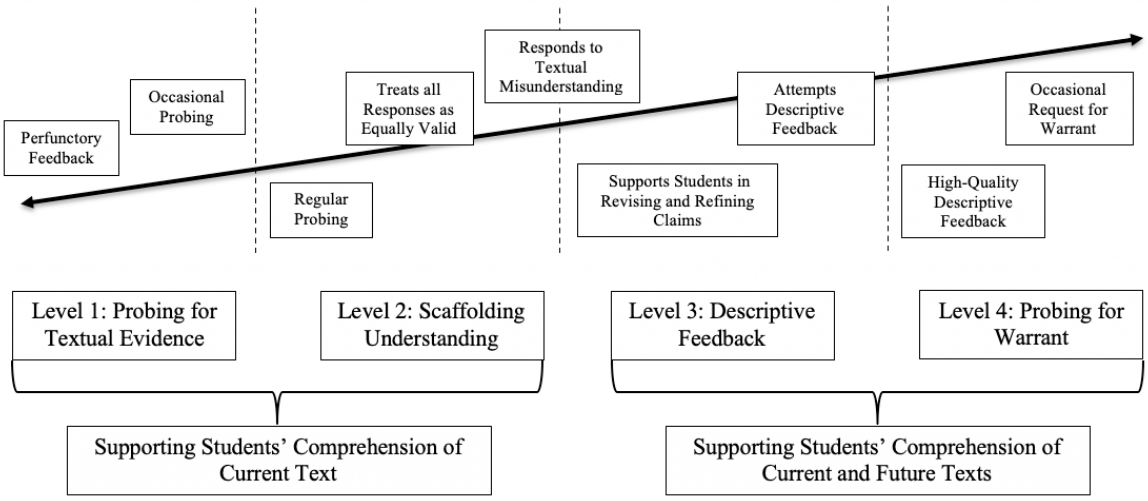
Source of Variation	Studies	Treatment effect	Magnitude of effects	Sign of effects	Significance patterns	No statistical difference between coaching effects	Estimated difference between coaching effects
Timing of study (Spring 2018 vs. Spring 2019)	Study 1 (N=99) *vs.* Study 3 (N=98)	1.65** (0.22) 1.37** (0.23)	✓	✓	✓	✓	-0.29 (*p*-value: .27)
Teaching task (Establishing norms vs. Providing feedback)	Study 2 (N=99) *vs.* Study 3 (N=98)	1.34** (0.25) 1.37** (0.23)	✓	✓	✓	✓	0.03 (*p*-value: 0.94[#])
Delivery (Online vs. In-person)	Study 3 (N=98) *vs.* Study 5 (N=104)	1.37** (0.23) 1.67** (0.20)	✓	✓	✓	✓	0.29 (*p*= 0.37)
Participant characteristics and concurrent coursework (Teacher candidates vs. undergraduates interested in teaching)	Study 3 (N=98) *vs.* Study 4 (N=95)	1.37** (0.23) 0.58 (0.36)	×	✓	×	×	-0.79* (*p*= 0.03)
	Adjusted Study 3 (N=98) *vs.* *Adjusted* Study 4 (N=95)	1.36** (0.23) 0.52 (0.37)	×	✓	×	×	-0.84* (*p*= 0.03)

Notes: [#] We use a bootstrapping procedure described by Steiner & Wong (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 4. *Adjusted* Study 3 and Study 4 includes treatment effect estimates obtained from regression

models that control for common baseline participant characteristics across the two studies. We obtained similar results for *adjusted* coaching effects when sample participants in Studies 3 and 4 were matched on baseline demographic characteristics using inverse propensity score weights. ✓ indicates that replication success was achieved, × indicates that replication failure occurred. 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 Text-Based Feedback Teaching Scenario



A2. 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

A3. Summary of Conceptual Replication Designs, Assumptions, and Diagnostic Results

The conceptual replication effort was designed according to the Causal Replication Framework introduced by Steiner, Wong, & Anglin (2019), and discussed in Wong, Anglin, & Steiner (2021). The Framework describes five sets of assumptions under which replication success can be expected. The assumptions may be understood broadly as *replication design* requirements (R1-R2), and *individual study design* requirements (A1-A3). Replication design assumptions include treatment and outcome stability (R1) and equivalence in causal estimands (R2) across studies. Combined, these two assumptions ensure that the same causal estimand for a well-defined treatment-control contrast, target population, and setting is targeted across all studies. Individual study assumptions address the causal identification of causal estimands (S1), unbiased estimation of causal estimands (S2), and correct reporting of estimands, estimators, and estimates (S3). These assumptions ensure that valid research designs are used for identifying study-specific effects, unbiased estimators are used for estimating effects, and effects are correctly reported (these are standard assumptions in most individual causal studies). When one or more of the replication and/or individual study assumptions are not met, direct replication of effects usually fails.

An advantage 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. *Direct replications* examine whether two or more studies with identical causal estimands yield the same effect (akin to the definition of verification tests in Clemens, 2017). This should be the case (within the limits of sampling uncertainty) if all causal replication and individual study assumptions are met. In contrast, *conceptual replications* examine whether two or more studies with intentionally varied causal estimands yield the same effect (akin to robustness tests proposed by Clemens, 2017). Here, the researcher introduces systematically planned violations in replication assumptions (R1-R2). For instance, a variation in treatment conditions, population characteristics, settings, or outcome measures. Conceptual replication designs include: *multi-site designs* with variations in participant or setting characteristics across sites (R2), *switching replication designs* with variations in settings across alternating intervention intervals (R2), *multiple cohorts* and *stepped-wedge designs* with variations in when treatments are introduced across time (R2), and *multi-arm treatment designs* with variations in treatment dosage levels (R1) (Wong et al., 2021). In each of these designs, if replication failure is observed, it is because of systematic differences in units, treatments, outcomes, settings, or time.

In this replication effort, we designed a series of conceptual replication designs to evaluate four systematic sources of effect variation – timing of the study, teaching task, delivery of simulation and coaching sessions, and target population and their concurrent coursework. Table A3 describes the replication design used (column 1), the replication and individual study assumptions examined (row 1), and the result of our diagnostics by comparing study characteristics on Table 1 for assessing the extent to which assumptions were met (✓ if we concluded that the assumption was met, × if we concluded that the assumption was not met).

Table A3. Conceptual Replication Design, Causal Replication Assumptions, and Results of Diagnostics for Addressing Assumptions

	R1. Treatment / Outcome Stability	R2. Equivalent Causal Estimand	S1. Identification	S2. Estimation	S3. Reporting
Timing: Multiple Cohort (Study 1 vs. Study 3)	Treatments ✓ Outcomes ✓	Participants ✓ Settings ✓ Causal quantity ✓ Time ×	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
Teaching Task: Switching Replication (Study 2 vs. Study 3)	Treatments ✓ Outcomes ✓	Participants ✓ Settings × Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓
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 ✓
Participant Characteristics & Concurrent Coursework: Conceptual Replication with Different Units and Settings (Study 3 vs Study 4)	Treatments ✓ Outcomes ✓	Participants × Settings × Causal quantity ✓ Time ✓	Balanced groups from the RCT ✓	Robust over multiple model specifications ✓	Verified by reanalysis from independent reporter ✓

A4. Assessing Implementation of Simulation Sessions and of the Coaching Protocol.

To assess *implementation of the coaching protocol*, the research team adopted a *natural language process* method introduced by Anglin, Wong, & Boguslav (2021) that uses semantic similarity methods to quantify the similarity between transcripts of intervention sessions. Anglin et al. demonstrates that these methods can be used to assess intervention fidelity in evaluation settings with highly standardized protocols that are delivered through verbal interactions with participants. In these cases, texts from transcripts of intervention sessions can be represented by their vocabulary and compared to one another by the relatively 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.

Adherence scale scores obtained from semantic similarity methods range from 0 to 1, where transcripts of coaching sessions with high adherence to the coaching protocol have higher scores, and those that stray from the protocol have 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). See Anglin et al. (2021) for further description of the method and interpretation of scores.

A5. 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 A6. 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 A7. 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 A8. 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 A9. 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.