ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 23-862

Different methods for assessing pre-service teachers' instruction: Why measures matter

Arielle Boguslav
Brown University

Julie Cohen
University of Virginia

Teacher preparation programs are increasingly expected to use data on pre-service teacher (PST) skills to drive program improvement and provide targeted supports. Observational ratings are especially vital, but also prone to measurement issues. Scores may be influenced by factors unrelated to PSTs' instructional skills, including rater standards and mentor teachers' skills. Yet we know little about how these measurement challenges play out in the PST context. Here we investigate the reliability and sensitivity of two observational measures. We find measures collected during student teaching are especially prone to measurement issues; only 3-4% of variation in scores reflects consistent differences between PSTs, while 9-17% of variation can be attributed to the mentors with whom they work. When high scores stem not from strong instructional skills, but instead from external circumstances, we cannot use them to make consequential decisions about PSTs' individual needs or readiness for independent teaching.

VERSION: October 2023

Suggested citation: Boguslav, Arielle, and Julie Cohen. (2023). Different methods for assessing pre-service teachers' instruction: Why measures matter. (EdWorkingPaper: 23-862). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/h94x-as50

Different methods for assessing pre-service teachers' instruction: Why measures matter

Arielle Boguslav and Julie Cohen

Abstract

Teacher preparation programs are increasingly expected to use data on pre-service teacher (PST) skills to drive program improvement and provide targeted supports. Observational ratings are especially vital, but also prone to measurement issues. Scores may be influenced by factors unrelated to PSTs' instructional skills, including rater standards and mentor teachers' skills. Yet we know little about how these measurement challenges play out in the PST context. Here we investigate the reliability and sensitivity of two observational measures. We find measures collected during student teaching are especially prone to measurement issues; only 3-4% of variation in scores reflects consistent differences between PSTs, while 9-17% of variation can be attributed to the mentors with whom they work. When high scores stem not from strong instructional skills, but instead from external circumstances, we cannot use them to make consequential decisions about PSTs' individual needs or readiness for independent teaching.

Over the last two decades, teacher preparation programs (TPPs) have become increasingly responsible for the collection and analysis of data on pre-service teachers' (PST) knowledge, skills, and dispositions (Bastian et al., 2016). Several states have implemented accountability systems that require TPPs to provide data on their graduates, share these data as a measure of program effectiveness, and even levy penalties on programs deemed ineffective (Levine, 2006; Texas State Legislature, 2009; US Department of Education, 2011). The theory of action behind these policies is TPPs can use PST data to provide more targeted support based on PSTs' needs, as well as adjust programmatic experiences based on the degree to which such experiences promote PSTs' development (Bastian et al., 2018; Davis & Peck, 2020).

Among other measures, most TPPs collect observational ratings of PSTs' instructional practice during clinical placements (Feuer et al., 2013). In capturing observable skills—what PSTs *do* in interactions with students—these measures are vital because TPPs are ultimately charged with preparing novices who engage in productive and supportive interactions with children. These measures are particularly challenging to implement in reliable ways, however (Bell et al., 2012; Gitomer, 2009; T. Kane & Staiger, 2012). Ratings of teaching practices are often influenced by factors that may not be related to a teacher's instructional skills, including rater standards and the characteristics of students in a classroom (Bartanen & Kwok, 2021; Campbell & Ronfeldt, 2018). As a result, TPPs risk erroneously attributing differences between external factors to differences in PSTs' instructional skills. For example, if observational scores are substantially influenced by supervisors' rating standards, the PSTs with the lowest scores may not be those most in need of targeted support, but instead, may be those rated by supervisors with harsher standards (Bartanen & Kwok, 2021). Moreover, unlike in-service teachers, PSTs are not solely responsible for their own classrooms, but instead teach in the context of mentors'

classrooms, whose characteristics and teaching skills may further influence PST scores on observational measures.

Figuring out how to measure instructional quality has been a longstanding puzzle for education researchers, but few have explored the challenges particular to doing so in the context of pre-service preparation, when time is in short supply and candidates are typically observed in contexts that are not fully "theirs." TPPs use PST observation scores to meet accreditation reporting requirements and guide programmatic decisions (AACTE, 2018; CAEP, 2022). If these measures suffer from the challenges described above, then we must re-evaluate what conclusions can reasonably be drawn from those data.

In this paper we investigate these measurement challenges using two distinct observational measures of PSTs' instructional skills, implemented as part of one TPP's efforts to generate robust data on such skills. The first measure is the Classroom Assessment Scoring System (Pianta & Hamre, 2009), designed to provide a holistic view of PSTs' instruction during clinical placements. The second measure provides a more focused view of PSTs' implementation of discrete instructional skills culled from simulation-based Instructional Activities (Lampert & Graziani, 2009).

In investigating these challenges, we focus on two key concepts: reliability and sensitivity. Following several prior studies of the reliability of observational measures, we define reliability according to the generalizability theory framework (Bartanen & Kwok, 2020; Briggs & Alzen, 2019). Under generalizability theory, a PST's score on an observational rubric consists of their "true score," the signal, and measurement error, the noise. Rather than treating measurement error as a single, monolithic construct, as classical test theory does, generalizability theory decomposes error into distinct sources. For PSTs, we conceptualize the signal as

information a TPP wants to learn about a PST's instructional skills. Crucially, because PSTs could go on to teach in a wide variety of contexts, TPPs are ultimately concerned with making inferences about individual PSTs' *transferable* skills rather than their skills in a particular classroom context working alongside particular mentor (Bell et al., 2012). To provide reliable information about PSTs' skills, observation scores must *generalize* beyond the context of the observation. In other words, under generalizability theory, a perfectly reliable measure will result in the same, stable score for an individual PST regardless of the context in which a PST is observed or the rater assessing their skills.

In practice, of course, no measure is perfectly reliable, which is why generalizability theory distinguishes stable signal about PSTs from sources of measurement error, thereby providing an estimate of how much of the variation in scores reflects signal we care about. Demonstrating reasonably strong reliability (a.k.a. generalizability) is a necessary but insufficient precursor for demonstrating predictive validity. If observation scores primarily reflect rater standards or classroom context, then we can hardly expect them to provide TPPs with predictive information about PSTs' instructional skills as teachers of record. The reliability analyses we conduct in this paper are concerned with understanding the extent to which individual PSTs' observation scores are stable or generalizable across different contexts and observation conditions.

Sensitivity is a practical challenge, often created by low reliability. We conceptualize sensitivity as the extent to which measures can detect statistically significant and practically meaningful differences when making comparisons between individual PSTs, groups of PSTs, or aspects of instructional quality (M. Allen & Coble, 2018). The less reliable a measure is, the less likely it is to distinguish among differences in PSTs' instructional skills, because those

differences may be drowned out by external factors that influence scores. To provide PSTs with targeted support, TEPs need measures that are sensitive enough to identify which PSTs may have less-developed skills.

Based on the measures themselves and how they were implemented, we have hypotheses about the reliability and sensitivity of CLASS scores and simulation-based scores. First, we hypothesize that both researcher-created measures have the potential to be more reliable than more typical observations conducted by clinical supervisors or mentors using home-grown rubrics, and thus better able to detect consistent differences among PSTs (Bartanen & Kwok, 2021; Bastian et al., 2018). Second, although CLASS may be more reliable than typical observational measures completed by individuals who know the PSTs, we hypothesize scores may be substantially influenced by external factors that do not reflect PSTs' instructional skills, including mentor teachers. Third, we hypothesize that raters will contribute minimally to measurement error for standardized simulation-based observations because of the standardized nature of these measures and the rating procedures employed. Fourth, we hypothesize the stronger reliability of CLASS and simulation scores (relative to commonly used measures) will contribute to greater sensitivity to differences between individual PSTs, groups of PSTs, their learning experiences, and instructional skills.

We use data collected by the TPP to test these hypotheses by answering the following research questions:

1. To what extent does each measure capture consistent differences between PSTs? (Reliability)
2. To what extent does variation in CLASS scores reflect differences between external factors, namely mentor teachers? (Reliability)

3. To what extent does variation in simulation scores reflect differences between external factors, namely raters? (Reliability)

4. How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction? (Sensitivity)

In this way, we can identify both the strengths of these measures and the ongoing challenges for learning about PST practice.

The pre-service period is both brief and formative. TPPs have a very short window to provide experiences and supports that position candidates for success as teachers of record. Despite the urgency of preparing skilled PSTs, we know very little about the extent to which the data TPPs currently collect can inform this work, or what changes might be necessary if they are not. This paper addresses this gap in two ways. First, we empirically explore the affordances and constraints of two measures designed to make inferences about PSTs' instructional skills. Second, we identify specific steps that TPPs should consider to enhance the reliability and sensitivity of the data they collect, as well as the tradeoffs such steps entail.

Background & Conceptual Framework

Why observe PSTs' instruction?

Many argue TPPs would benefit from having more insight about the degree to which preparation experiences support PSTs' development (e.g., Davis & Peck, 2020; Goldhaber, 2019). In our conceptual framework (Figure 1), we articulate a theory of action for how data can result in improvements in teaching when PSTs enter the classroom. In the first column, we highlight four mechanisms by which PST data may support TPP improvement. In the remaining three columns, we indicate how TPPs may implement these mechanisms and highlight the proximal and distal outcomes.

The first mechanism highlights how measurement tools can facilitate a shared understanding and common language among TPP faculty, supervisors, mentors, and PSTs (Davis & Peck, 2020). This facilitates teacher preparation stakeholders working toward the same learning goals and coherence across preparation experiences.

The remaining three mechanisms highlight how specific analyses can improve the supports and experiences TPPs provide. First, analyzing PST data can allow TPPs to diagnose and respond to PSTs' individual learning needs (M. Allen & Coble, 2018; Bastian et al., 2018). Second, sufficiently sensitive PST data can help TPPs identify program-wide areas for development, which can inform changes to course content and program curriculum (M. Allen & Coble, 2018; Peck & McDonald, 2013). By comparing PSTs' mastery of different facets of instruction (e.g., relative strengths in classroom management versus facilitating discussions), TPPs can tailor learning experiences to meet observed needs (e.g., practice facilitating student discourse). Moreover, TPPs can compare skill development across cohorts or licensure tracks (e.g., elementary versus secondary) to identify areas of relative weakness and redesign or supplement program content in those areas. Finally, a TPP can compare the scores of PSTs that participated in different preparation experiences, such as comparing scores between PSTs who completed rehearsals of core teaching practices and PSTs who did not (McDonald et al., 2013). In this way, TPPs can identify experiences that are more promising for supporting PSTs' development (Hill et al., 2020; Peck & McDonald, 2013).

What makes a measure useful?

Realizing improvements from PST data depends on what data can tell us. Here we highlight two key measurement properties that influence the conclusions TPPs can draw: reliability, and sensitivity.

Reliability. In line with generalizability theory, we conceptualize reliability as the extent to which differences in scores reflect differences between PSTs (Bell et al., 2012; Ho & Kane, 2013). When reliability is low, differences in scores are influenced primarily by differences in the conditions of assessment, rather than differences between PSTs. Measures requiring human judgment, including observational measures, tend to be less reliable than paper and pencil tests with one correct answer (Bell et al., 2019; Hill et al., 2012). This is because it is very difficult to ensure raters assign scores in the same way (Bell et al., 2015). Furthermore, prior work suggests lesson content, time of year, and student demographics contribute to relatively low reliability of in-service teacher observation scores (Bell et al., 2012; Casabianca et al., 2015; Ho & Kane, 2013).

There is limited research on the reliability of PST observation scores. Bartanen and colleagues (2021) applied a generalizability framework to decompose variation in PST observation scores from a large TPP in Texas, where PSTs were rated by clinical supervisors. Typical to many TPPs in the US, supervisors were primarily tasked with supporting PST development but were also asked to rate PSTs at four timepoints during student teaching, using a proprietary observation rubric adapted from the state's teacher evaluation framework. The authors found only 20% of the variation in PST scores reflected consistent differences between PSTs, whereas 55% of the variation in scores could be attributed to differences in the stringency of supervisors' rating standards. The authors also suggest the potential for mentor effects: since PSTs are typically observed in their mentor's classroom, the classroom climate may be shaped by the mentor's instructional practices rather than those of the PST. To our knowledge, all other existing work on the reliability of PST data focuses on non-observational measures or on rater agreement, without considering other sources of statistical noise (e.g., Bastian et al., 2018).

The literature highlights two primary strategies for increasing measure reliability. First, TPPs can reduce the influence of assessment conditions by standardizing them. This includes providing raters with extensive training and feedback to ensure consistency, providing PSTs with a standardized lesson plan or learning objective, and observing PST instruction in the context of standardized teaching simulations (Cohen & Goldhaber, 2016; Cohen et al., 2020). Second, TPPs can isolate the differences between PSTs that are consistent across observations by averaging PST scores over multiple observations under a variety of conditions (T. Kane & Staiger, 2012). In this case, randomly assigning conditions, such as raters or lesson objectives, for each observation is especially helpful (van der Lans et al., 2016).

Sensitivity. Although sensitivity is a less developed empirical concept in education research than reliability, it connotes the degree to which a measure can detect differences a TPP might want to understand, such as which PSTs are most skilled, if some groups of PSTs are comparably more skilled, or if some aspects of instruction are comparatively stronger across all PSTs in a TPP (M. Allen & Coble, 2018; Hill et al., 2020; Mancenido, 2022). Data-driven decision-making is unlikely to lead to improvements in PSTs' preparedness if measures are not sufficiently sensitive to these differences. Identifying which PSTs need additional support implicitly requires comparing PSTs' scores. Understanding the impacts of specific preparation experiences requires comparing scores between the group of PSTs that completed an experience and those who did not. Similarly, identifying specific instructional skills for which PSTs need more support requires comparing scores from one skill to another. To accurately interpret these comparisons, the measures TPPs use must be sensitive enough to detect these differences (Cohen & Goldhaber, 2016; Weisberg et al., 2009). When measures are not sufficiently sensitive, TPPs

risk concluding that there are no differences, when in fact the measures may be unable to identify them.

Measure sensitivity can be influenced by a variety of factors. First, the less reliable a measure is, the less sensitive it is likely to be. When only a small portion of the variation in scores reflects consistent differences between PSTs, then these differences can be drowned out by the other external factors (raters, lesson context, etc.) that influence scores. Scoring procedures and the range of scores available also influence sensitivity. A measure with only four possible scores provides coarser differentiation among PSTs' skills than a measure with ten possible scores, assuming scores are reliably measured and PSTs receive the full range of scores (T. Kane & Staiger, 2012). Unfortunately, evidence suggests PSTs generally do not receive the full range of scores, with most scores clustering at the highest end (Kraft & Gilmour, 2017).

The scope and granularity of a measure can also influence sensitivity (Janssen et al., 2015). Measures that provide a broader picture of PSTs' skills, like "instructional support," tend to provide less detailed areas of strength or improvement than a fine-grained measure, such as the specific instances where PSTs engage in "feedback loops" with students (Hill et al., 2020; Mancenido, 2022). Because of reliability issues and logistical constraints, providing finer-grained information necessarily requires trading off breadth for depth. It is not feasible, for example, to ask raters to individually score hundreds of finer-grained indicators of PST skill (Bell et al., 2015). If the main differences between PSTs, however, are more nuanced, for example based on the frequency and substance of their feedback loops, then broader measures of instructional support will be less sensitive because they encompass much more than granular measures.

Common measures of PSTs' instruction are likely to be limited in sensitivity because of these issues. In addition to suffering from low reliability, these measures generally provide a broad picture of PSTs' skills, providing little information about finer-grained details of PST development (Hill et al., 2020; Mancenido, 2022). Additionally, these measures often use a limited score range of three to five points (Bartanen & Kwok, 2021; Henry et al., 2013). We are not aware of any literature that focuses on the sensitivity of PST observation scores. From Bartanen & Kwok's (2021) analyses, however, we can see that the difference in baseline scores and growth for a PST in the 16th percentile and 84th percentile (a 2 SD difference) is statistically significant, suggesting some ability to detect differences between PSTs. Bastian et al.'s (2018) identification of four distinct PST profiles based on edTPA scores also suggests some ability to detect differences between PSTs, although we note edTPA is a portfolio assessment not an observational measure.

Data & Methods

Context

We utilize data from a TPP at a large, public university in the southeastern United States, which we call Lambeth University. Lambeth offers multiple pathways for PST licensure, enrolling approximately 120 PSTs each year. PSTs at Lambeth are typically white, speak English as a first language, under 27 years old, and attended high schools with students of middle or high socio-economic status (Table A3).

Lambeth takes a practice-based approach to preparing PSTs (Forzani, 2014). Coursework emphasizes the development of robust content and pedagogical knowledge, while explicitly linking theory to pedagogies of practice (Grossman et al., 2009). In addition to role-play exercises, many instructors provide opportunities for PSTs to teach digital student avatars in

mixed-reality simulations, where PSTs can practice specific skills, techniques, and approaches before working with real children (Cohen et al., 2020; Dieker et al., 2014). This allows faculty to observe and assess PSTs' teaching in ways that are otherwise difficult to replicate in a university classroom.

PSTs also complete clinical experiences each semester focused on application of course content. Early clinical experiences vary by program but range from one-on-one reading tutoring to interning in a classroom for 15 hours a week. Across programs, the final semester features a full-time student teaching placement designed to afford experience with all aspects of teaching. During each clinical experience, PSTs complete multiple coaching cycles, following a modified version of the evidence-based *My Teaching Partner* (MTP) program (J. Allen et al., 2015). In each cycle, PSTs record a lesson in their clinical placement. Field supervisors analyze video segments and provide PSTs with reflective prompts before meeting with the PST to discuss PSTs' strengths and areas for growth.

Measures

Lambeth dedicated substantial resources to a robust data collection system for the purposes of accreditation, research, and program improvement. Lambeth systematically collects data on PST instructional practice at multiple timepoints using the Classroom Assessment Scoring System (CLASS), designed to capture broad features of classroom climate and instructional support in their placement (Pianta & Hamre, 2009), and finer-grained measures of teaching in standardized mixed-reality simulations (Cohen et al., 2020). Lambeth uses CLASS data for accreditation purposes and to research PSTs' preparedness to teach. Unlike the CLASS, the simulation measures were designed to identify differences between PSTs who received instructional coaching between simulation sessions and those who self-reflected to discern which

method of scaffolding better supported PSTs' skill development (Cohen et al., 2020). To date, such measures have not been used to compare individual PSTs' skills. However, Lambeth is interested in using these measures moving forward to answer questions about PSTs' instructional skills, including growth over time, differences between groups of PSTs who have different learning experiences (e.g., courses with different instructors), and program-wide strengths and weaknesses.

For both measures, the university employs rating procedures that are aligned with "best practice" in the classroom observation literature. Raters have no relationship with PSTs they rate. They also complete formal trainings, are required to pass a rater certification test, and receive ongoing feedback at weekly scoring calibration meetings (Park et al., 2015). Finally, observation videos are randomly assigned to different raters at each observation timepoint (T. Kane & Staiger, 2012).

CLASS Scores

PSTs select up to four videos submitted for MTP coaching for external scoring. Videos are independently scored by Lambeth using the CLASS framework (Pianta & Hamre, 2009), an observation protocol emphasizing the value of positive relationships, focusing on the tenor of interactions between a teacher and students and among students, and qeating the *classroom* as the unit of analysis. To date, the framework has been used for measuring classroom quality for research purposes, guiding coaching conversations, and evaluating the quality of early childcare programs (J. Allen et al., 2015; Araujo et al., 2016; Bassok et al., 2021). CLASS provides a high-level view of three broad domains: Emotional Support, Classroom Organization, and Instructional Support, each comprised of 3-5 dimensions (Hafen et al., 2015). For example, one dimension within the Classroom Organization domain is Productivity. One dimension within

Instructional Support is Quality of Feedback. Raters score each dimension on a 7-point scale, using dimension-specific indicators to distinguish between teacher-student and student-student interactions that are low (1-2 points), mid (3-5 points), or high quality (6-7 points).

Prior research documents the reliability, sensitivity, and validity of CLASS when used to assess classroom quality for in-service teachers. When scores from multiple observations—each rated by a different rater—are averaged together, over 60% of the variation in scores can be attributed to differences between teachers (T. Kane & Staiger, 2012). CLASS scores have also been used to detect differences between teachers and document changes in teaching over time (Bassok et al., 2021; T. Kane & Staiger, 2012; La Paro et al., 2004). Finally, higher CLASS scores are associated with other measures of classroom quality and stronger student outcomes, including academic performance and student engagement (e.g., J. Allen et al., 2013; Araujo et al., 2016; T. Kane & Staiger, 2012; La Paro et al., 2004). However, a growing body of literature is raising concerns about the influence of contextual factors on CLASS scores and our ability to use these scores to explore change in teacher practice over time (reliability) and to detect differences between teachers (sensitivity) (Briggs & Alzen, 2019; Casabianca et al., 2015; Gitomer et al., 2014; Wallace et al., 2020).

Simulation-Based Scores

Beginning in 2017-18, all Lambeth PSTs also participated in two standardized simulations as part of their general methods courses. In these simulations, PSTs teach virtual students voiced by a trained actor, providing an opportunity to engage in "approximations of practice" (Grossman et al., 2009). The simulations also serve as a standardized platform to assess skill development. In the first simulation scenario, referred to as "Redirection," PSTs practiced redirecting standardized off-task student behaviors in a whole-class discussion (Cohen et al.,

2020). In the second simulation scenario, referred to as "Text-Focused Instruction," PSTs facilitated a discussion and responded to standardized student responses that were supported by textual evidence to different degrees (Cohen et al., 2023). PSTs completed each simulation scenario four times over the course of the program, though with different interval spacing than CLASS observations. Figure 2 illustrates the timeline for each simulation-based observation and when PSTs submitted videos for CLASS scoring.

The measures used to assess PSTs' instructional skill in the simulations are specific to each scenario. The rubric for the Redirection simulation was based on the Responsive Classroom framework that local K-12 schools used (Charney, 1993; Responsive Classroom, 2014). The rubric for the Text-Focused Instruction simulation was designed to reflect high-quality instructional practices in the relevant literature (Castles et al., 2018; Reznitskaya et al., 2009). The simulation rubrics assess more granular practices than the CLASS. Whereas the Redirection simulation rubric focuses entirely on behavioral redirection, for example, CLASS treats "effective redirection of misbehavior" as a component of the dimension of Behavior Management, which is, in turn, a component of the Classroom Organization domain. Nevertheless, there is conceptual overlap between the granular simulation rubrics and broad CLASS rubrics.

From a measurement standpoint, both the CLASS and simulation scores are criterion-referenced rubrics that require raters to select a specific score along a continuum from low to high. CLASS uses a slightly restricted score range of 1-7 compared with the simulation score range of 1-10. Additionally, while raters select a single score for each simulation rubric, scores for Classroom Organization and Instructional Support are generated as an average of rater scores across multiple sub-dimensions.

As part of a larger research study, PSTs were randomly assigned to a short coaching session or a short self-reflection protocol in between simulation sessions. Roughly half the PSTs received coaching for the Redirection scenario and the other half for the Text-Focused Instruction scenario. For more detailed descriptions of the simulation measures and protocols, see Author 2020, 2021.

Sample

For each measure, we selected a sample that maximized sample size, while also ensuring a similar number and timing of observations, to avoid bias from missing data. For CLASS scores, our sample consists of the 83 PSTs who entered Lambeth in 2017-18 and for whom four observations are available, two from each of two clinical placements. In total, this sample includes 135 mentors[1] from 62 schools across more than 10 counties in one state. Schools are primarily elementary schools serving mostly students who are white and not eligible for free or reduced-price meals. For simulation scores, our sample consists of the 60 PSTs who entered Lambeth in 2018-19 and for whom eight observations (four for each simulation scenario) are available. Appendix A includes a detailed discussion of the full dataset and rationale for sample restrictions. Though the samples for each measure are different, a covariate balance test (Table A3) suggests that there are few demographic differences between PSTs from the 2017-18 cohort (CLASS analyses) and PSTs from the 2018-19 cohort (simulation analyses).

Methods

Given our focus on cross-measure comparisons, we analyze CLASS scores for the two domains most closely aligned to the constructs measured by simulation rubrics: Classroom Organization (aligned with Redirection) and Instructional Support (aligned with Text-Focused

[1] Our sample includes more mentors than PSTs, because PSTs work with one mentor during early clinical experiences and a different mentor for their formal student teaching experience.

Instruction). For each research question, we conduct analyses separately for each simulation scenario and CLASS domain. Below we provide a general overview of our methods. Detailed explanations of the statistical models for each research question are included in Appendix B.

RQ1: To what extent does each measure capture consistent differences between PSTs?

We draw on generalizability theory (Bartanen & Kwok, 2020; Briggs & Alzen, 2019) to decompose variation in observation scores into distinct sources, with the goal of distinguishing between: 1) variation that reflects consistent differences between PSTs, and 2) measurement error that results from differences in the conditions and context of measurement. To do so, we estimate PST observation scores as a function of the number of months since the first observation using a multi-level model with PST random effects. We then calculate the proportion of the variation that reflects consistent differences, relative to the overall variation in scores.

RQ2&3: To what extent does variation in scores reflect differences between raters and mentors?

We decompose variation in observation scores into contextual sources we hypothesize might influence scores without contributing to our understanding of PSTs' instruction. For CLASS scores, we separate consistent differences between PSTs from variation between mentors and any remaining measurement error. For simulation scores, we separate consistent differences between PSTs from variation between raters and any remaining measurement error. To do so, we augment the multi-level models employed in RQ1 to include mentor random effects (for CLASS scores) or rater random effects (for simulation scores).

RQ4: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

To explore the sensitivity of CLASS and simulation scores to individual differences, we review the estimates from the multi-level models used in RQ1with the aim of understanding the extent to which we can detect between-PST variation in scores and trajectories. Specifically, we use log likelihood tests to compare models with and without PST intercept and slope random effects to determine whether there is significant between-PST variation in baseline scores and growth trajectories. To explore the sensitivity of scores to group differences, we review results from RQ1 for simulation scores, where a control for participation in coaching or self-reflection allows us to evaluate whether simulation scores are sufficiently sensitive to detect differences between PSTs exposed to these preparation experiences. To explore sensitivity to different facets of instruction, we graphically compare scores on the two CLASS domains and scores for the two simulation-based measures. We also estimate the correlations between scores, comparing the two CLASS domains (Classroom Organization and Instructional Support) to each other and the two simulation-based measures to each other (Redirection and Text-Focused Instruction).

Limitations

Lambeth is not representative of all TPPs. In particular, Lambeth PSTs are largely white, female, and have college educated parents. We see our analyses as proof of measurement issues that can surface across a range of TPP contexts, but we are not making any claims here about the degree to which these findings are generalizable. There are also several limitations of the measures we employ, which necessarily reflect specific conceptualizations of core components of high-quality instruction. Many crucial aspects of teaching, including culturally and linguistically responsive and sustaining pedagogies, are not included in these measures (Pacheo, 2009; Paris & Alim, 2014). We also do not yet know the extent to which CLASS scores from clinical placements or simulation-based scores provide meaningful information about PSTs'

instruction when they become teachers of record. Though CLASS scores have been used to evaluate PSTs' instructional skills during student teaching (e.g., Malmberg et al., 2010), we are not aware of any studies that evaluate the predictive validity of CLASS scores for instructional skills down the road. Such evidence is also lacking for simulation scores, though such research is currently under way (Cohen et al., 2023). Unfortunately, generating this evidence using data from Lambeth is virtually impossible because of the lack of state-wide longitudinal data systems that connect PSTs to their later employment and teaching outcomes. Nonetheless, these kinds of predictive validity analyses are only possible when measures primarily reflect differences between PSTs. When measures are heavily influenced by contextual characteristics, such as raters or mentors, they cannot tell us much about differences between PSTs' skills, let alone predict what PSTs will do when they begin full-time teaching. We therefore argue that our analyses raise important considerations for TPPs and provide a proof-of-concept relevant to any measure of PSTs' skills, including those that capture other aspects of teaching.

Our analyses are also limited by unavoidable deviations from the ideal design for a generalizability study, which requires each PST be observed by every rater, in every kind of classroom context, and teaching every kind of lesson (Briggs & Alzen, 2019). Under these conditions, we could directly measure how each contextual characteristic influences scores. Random assignment of rater, classroom context, and lesson provide a more feasible alternative to estimate the average effects of each contextual characteristic. Because mentors are not randomly assigned to PSTs in our data, effects attributed to mentors may instead reflect differences between PSTs. This would be the case if PSTs with weaker instructional skills tend to be intentionally assigned to more skilled mentors. The estimated between-mentor variation in scores

would then reflect both these initial differences in PSTs' instructional skills *and* the influence of mentors on PST scores.

There are two ways mentors may influence PST scores, each of which has different implications for the interpretation of our results. First, mentors may influence PST scores by supporting PST skill development. Indeed, a growing body of literature highlights the important role mentors play in supporting PSTs' instructional skills as teachers of record and feelings of preparedness (Bastian et al., 2022, 2023; Goldhaber et al., 2020a, 2020b; Ronfeldt et al., 2018, 2020, 2021). In this case, differences between mentors' ability to support PST skill development would be included in our estimates of between-mentor variation in scores, even though they conceptually reflect "true" differences between PSTs that TPPs would want to discern. At the same time, mentors may also influence PST scores in ways that do not reflect PSTs' instructional skills. This would be the case, for example, if PSTs earn higher Classroom Organization scores when observed in classrooms taught by mentors with robust classroom routines the PST neither established nor maintained. Conceptually, we would want to remove this kind of variation in scores before drawing conclusions about between-PST differences in management skills. Unfortunately, we are unable to separate these two kinds of mentor influences in our data. One way this has been done in prior work involves estimating reliability separately at different timepoints to capitalize on the expectation that mentors' developmental effects increase over time, the desired "signal," while their indirect effects on PSTs' scores, the contextual "noise," would likely decrease as PSTs take over more responsibility in the classroom (Bartanen & Kwok, 2021). However, our data do not allow us to estimate reliability for individual timepoints

or separately by semester.[2] Rather than providing precise estimates of the proportion of variation in CLASS scores attributable to true differences in PSTs instructional skills, our results provide an initial estimated range of variation, while clearly problematizing the use of raw scores in drawing inferences about differences in PST skill.

Our analyses are also limited by the lack of access to data on all contextual characteristics of interest. While we have access to rater information for simulation scores, we do not have access to rater information for CLASS scores. Any attempt to decompose the variation in CLASS scores, therefore, suffers from the problem that not all potential contextual factors are accounted for in the model. This means that our estimates serve as an upper bound, as the true proportions would be lower if any unobserved factors influence scores (Briggs & Alzen, 2019).

Finally, we acknowledge several challenges with making direct comparisons between CLASS and simulation results. First, we drew on different samples, raising the possibility that observed differences in the scores' measurement properties could stem from sample rather than measure differences. However, we see little evidence of systematic differences between the samples on observable characteristics (Table A3). Moreover, the rating and scoring procedures were highly standardized and consistent across both AY2017-18 and AY2018-19. Additionally, we cannot rule out other potential differences between simulation and CLASS scores including characteristics of the rubrics themselves or unknown differences in the implementation of scoring procedures, including CLASS raters, that could theoretically influence their measurement properties. However, our goal is not to make conclusive claims about the cause of

[2] We cannot estimate reliability at a single point in time because most mentor teachers only work with one PST at a single timepoint, so we cannot separate out between-mentor variation. We can only estimate between-mentor variation in scores when using data from both semesters together, because PSTs change mentors from the first semester to the second. However, this doesn't allow us to estimate reliability separately by semester.

measurement issues we observe. Instead, we are surfacing factors that are within TPPs' control and that prior literature has suggested might play a role, including raters and mentors.

Results

RQ1: To what extent does each measure capture consistent differences between PSTs?

The proportion of variation in CLASS scores that reflects differences between PSTs is low relative to the overall variation in scores. Specifically, we estimate that 3-4% of the variation in individual Instructional Support and Classroom Organization scores reflects consistent differences between PSTs (Table 1). This means that 96-97% of the variation in scores reflects measurement error. Figure 3 shows the remaining variation in scores that reflects differences between PSTs, after we account for measurement error. It is very limited, especially for Classroom Organization scores.[3] In the second row of Figure 3, we plot the growth trajectories for all PSTs to illustrate how baseline scores and growth vary over time. Here, the y-axis reflects each PSTs' simulation or CLASS score, and the x-axis reflects the observation timepoint. We find 15% of the variation in growth in Instructional Support scores reflects consistent differences between PSTs over time (Table 1). The estimate for growth is higher because PST growth is calculated using all four scores, each of which contains signal of PST skill, while the proportion for individual scores is calculated only a single score.

Considering the many strengths of the CLASS measure and our first hypothesis that CLASS scores would be more reliable than the home-grown measures documented in prior literature, our results are surprising and dismaying. Bartanen & Kwok (2021), for example, found that 20% of the variation in scores reflected consistent differences between PSTs even

[3] Figure 3 also illustrates some surprising trends in PSTs' instructional practice over time. In particular, Classroom Organization scores appear to decrease over time. Unfortunately, we do not know why this is the case. We also see evidence that PSTs' Instructional Support scores dip between the spring of their first year and the beginning of the second year, which we hypothesize might result from the summer break away from the classroom.

though their data did not rely on an established observational rubric and scores were generated by supervisors that knew the PSTs they rated and received much less robust rater training.

We also find a greater proportion of the variation in simulation scores reflects consistent differences between PSTs. Specifically, about 20% of the variation in scores reflects consistent differences between PSTs, as compared with the 3-4% for CLASS scores (Table 1), about a five-fold difference. When compared with the prior literature, these results are somewhat more encouraging. In addition to being consistent with Bartanen & Kwok's (2021) findings, these results are consistent with prior work on the reliability of in-service teacher observation scores under conditions more favorable for reliability[4] (Briggs & Alzen, 2019; Ho & Kane, 2013; T. Kane & Staiger, 2012).

The estimates for variation in growth over time are similar for CLASS and simulation scores (Table 1). Directly comparing these two estimates is misleading, however, since the growth estimates are scaled by the intervals between observations, which are far greater for CLASS scores (see Appendix B). To allow more direct comparison, we analyze the proportion of variation in growth if CLASS scores and simulation scores were collected at the same intervals, by recalculating the estimate for simulation scores using the scaling factor for CLASS scores. If simulation scores were separated by the same intervals as CLASS scores, about 40% of the variation in growth on Text-Focused Instruction scores would reflect consistent differences between PSTs. This is about two and half times larger than the estimate for CLASS Instructional Support.

[4] The data used included double the number of observations per teacher, and closer spacing between observations. Furthermore, because the data come from observations of in-service teachers there are no concerns about measurement error stemming from mentor teachers.

RQ2: To what extent does variation in CLASS scores reflect differences between external factors, namely mentor teachers?

After separating out between-mentor score variation, 9-17% of the variation in CLASS scores can be explained by the mentor in whose classroom a PST is observed (Table 2).[5] Once we account for variation between mentors, the proportion of variation that reflects differences between PSTs falls to effectively zero. These results differ from Bartanen & Kwok's (2021) analysis, where they find little evidence mentors contributed to variation in scores but are consistent with our second hypothesis that CLASS measures would likely capture mentor effects.

A key question here is how to interpret our results. Some between-mentor variation may reflect differences in the classroom context (e.g., classroom routines previously established by the mentor) unrelated to differences in PSTs' instructional skills. At the same time, mentors may also differ in their ability to support PST skill development, in which case between-mentor variation in scores may reflect true differences in PST skill. Additionally, because mentors were not randomly assigned to PSTs, between-mentor variation may also reflect non-random sorting of more skilled PSTs to certain mentors. In the latter two cases, removing between-mentor variation in scores would understate the true differences between PSTs.

While we cannot fully disentangle these sources of between-mentor score variation, we can engage in a bounding exercise. In the RQ1 models, where we don't account for between-mentor variation, any variation in scores stemming from mentors' developmental effects or PST sorting are absorbed as part of the estimated between-PST variation in scores. The estimates

[5] Because of challenges estimating the proportion of variation in scores between mentors for Classroom Organization, we also estimate an alternative model where we only consider differences between mentors, ignoring any additional variation that stems from differences between PSTs. Under this model, we find that 15% of the variation in Instructional Support scores reflects consistent differences between mentors. Similarly, we find that 17% of the variation in Classroom Organization scores reflects consistent differences between mentors.

from Table 1, therefore, provide an upper bound for how much variation in PST scores may stem from PST sorting, mentors' developmental effects, and other differences in PSTs' instructional skills. When we decompose variation in scores into only between-mentor variation and measurement error, however, we find that 15-17% of the variation in scores can be explained by mentors alone, more than three times the upper bound provided by the Table 1 estimates (3-4%). This discrepancy suggests at least some of variation in raw PST scores stems from differences between the mentor context, above and beyond the contributions from PST sorting, mentor developmental effects, and PSTs' underlying differences in skill. In other words, observing PSTs in different mentor classrooms appears to contribute to the large proportion of the variation in CLASS scores indicated as measurement error (96-97%) in our initial RQ1 analyses.

RQ3: To what extent does variation in simulation scores reflect differences in external factors, namely raters?

Unlike mentors, who are not randomly assigned to candidates, raters *were* randomly assigned to observe specific PSTs at specific timepoints. Thus, we are confident that between-rater variation in scores does not result from systematic sorting of PSTs to raters (T. Kaine & Staiger, 2012). We also have no reason to believe that raters, who do not know or interact with these PSTs, would have any developmental effects on PSTs' instructional skills like the supervisors who often score PSTs (Bartanen & Kwok, 2021). Therefore, we can interpret any observed between-rater variation in scores as stemming from differences in rater standards, which TPPs would want to account for before making inferences about PSTs' skills. In line with our third hypothesis, raters explain only 1-3% of score variation (Table 2) after separating out between-rater variation in scores. This suggests differences in simulation scores between PSTs

are not substantially influenced by differences in rater standards, in contrast to the influence of supervisors on PSTs' ratings in Bartanen & Kwok's (2021) work.

RQ4: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

Sensitivity to Differences between PSTs

The lower the proportion of variation in scores that reflects differences between PSTs, the less likely it is for a measure to be sensitive enough to detect those differences. Our fourth hypothesis was the hoped-for higher reliability of CLASS and simulation scores would correspond with enhanced sensitivity of those measures. Though reliability for neither measure was ideal, we return to Figure 3 to understand the differences that CLASS and simulation scores *can* detect between PSTs, isolating only between-PST differences after accounting for measurement error, which is the most conservative approach. Results from additional analyses to quantify these differences and evaluate their statistical significance are included in Appendix D.

Differences between PSTs are small for CLASS scores. The difference in baseline CLASS scores between a PST at the 16th percentile and a PST at the 84th percentile corresponds to 0.2-0.3 points (out of 7) for both Instructional Support and Classroom Organization (Table D1). This difference is only statistically significant for Instructional Support. There is also a statistically significant difference in growth rate for Instructional Support, corresponding to a difference of 0.03 points between a PST at the 16th percentile and a PST at the 84th percentile. Relative to the 7-point CLASS scale, these differences represent at most 5% of the maximum possible difference between PSTs.

Consistent with the greater reliability of simulation scores, differences between PSTs on simulation scores are statistically significant and larger. The difference in baseline simulation

scores between a PST at the 16th percentile and 84th percentile is 1.11 points for Text-Focused Instruction and 1.71 for Redirection (Table D2). Relative to the 10-point simulation scale, this represents 12-19% of the maximum possible difference in scores, more than double the estimate for CLASS scores. These differences are similar in magnitude to the differences Bastian et al. (2018) observed, on average, between PSTs on the edTPA.[6] We acknowledge the possibility these results stem from PSTs having similar instructional skills, rather than a lack of measurement sensitivity. However, anecdotal evidence from Lambeth teacher educators and prior work documenting large differences between PSTs once they enter the classroom, suggest this is unlikely to be the case (e.g., Atteberry et al., 2015; Boyd et al., 2008).

Sensitivity to Differences between Groups-Simulation Scores

To explore score measurement sensitivity to different learning experiences, we compare simulation scores between PSTs who participated in coaching versus self-reflection between practice sessions. PSTs who received coaching for the Text-Focused Instruction simulation score 1.5 points higher immediately after coaching and 0.5 points higher when observed 5 months later, though the latter estimate is not significant in some models (Table C3). PSTs who received coaching for the Redirection simulation score 2.5 points higher immediately after coaching and 1.3 points higher two months later.

These results highlight the simulation-based measures are sensitive enough to detect differences between PSTs who participated in different preparation experiences immediately following such experiences and over time. Additionally, these results reinforce the feasibility of comparing groups of PSTs, rather than individual PSTs, even when a large portion of the variation in scores reflects measurement error. If measurement error and the influence of

[6] Unfortunately, it is difficult to compare our results to Bartanen & Kwok's (2021) because they use standardized scores.

contextual characteristics are the same across groups of PSTs, we can safely make comparisons across groups.

Sensitivity to Differences between Facets of Instruction- CLASS & Simulation Scores

Here, we compare PSTs' scores between Classroom Organization and Instructional Support to explore sensitivity to differences between facets of instruction. In Figure 4, using predicted scores to account for measurement error, we graph scores over time for all PSTs to see if they perform similarly across CLASS domains. Across both graphs, Classroom Organization scores are consistently higher than Instructional Support scores, with a difference of 1.00-3.25 points out of 7. This suggests PSTs in our sample are, on average, considerably stronger in management skills than providing instructional support. These results provide suggestive evidence that CLASS scores are sufficiently sensitive to identify program-wide patterns in areas of relative strength.

Table 3 shows that scores between domains are moderately correlated with one another (0.20-0.45). This means PSTs that receive higher scores on Instructional Support also tend to receive higher scores on Classroom Organization. This does not necessarily contradict the previous findings. However, it raises the possibility that there may be differences between PSTs that CLASS scores are not able to detect. This would be the case, for example, if raters perceive Instructional Support as more challenging—resulting in lower average scores—but also form a general impression of each PSTs overall skills (i.e., halo effects), rather than considering each domain individually (Cohen & Goldhaber, 2016). Under these conditions, PSTs with relatively higher Instructional Support scores could potentially be given high Classroom Organization scores, even if their underlying management skills were weaker. Unfortunately, we cannot

determine if this is the case at Lambeth. It is equally possible that PSTs at Lambeth simply have similar relative strengths and areas of improvement.

Results for simulation scores follow the opposite pattern. In Figure 5, we graph Redirection and Text-Focused Instruction scores over time for all PSTs and see more overlap in scores across the two facets of instruction within each group of PSTs (coaching vs. self-reflection). This suggests simulation scores may not be very helpful in understanding program-wide patterns in PSTs' relative strengths and areas of improvement. At the same time, correlations between Redirection and Text-Focused Instruction scores are quite weak (Table 3), providing empirical evidence that each measure captures a different facet of PSTs' instructional skills and suggesting simulation scores can help illuminate relative strengths and areas of improvement for individual PSTs.

Discussion and Implications

In many ways, Lambeth University is at the forefront of measuring pre-service teachers' instructional skills. Instead of relying on problematic supervisor ratings (Bartanen & Kwok, 2021), the university employs trained and certified raters who do not know the PSTs they rate, resulting in high inter-rater reliability. Furthermore, raters are randomly assigned to videos to avoid systematic bias in scores from differences in rater standards. The university also draws on a validated observation protocol (CLASS) and innovative researcher-developed, simulation-based measures that allow the university to standardize the lesson context in which a PST is observed. The university's data collection procedures are well-aligned with established best practices for in-service teacher observations (Ho & Kane, 2013), representing substantial improvements over typical approaches to measuring PST practice (Bartanen & Kwok, 2021; Mancenido, 2022).

Nevertheless, our results raise serious concerns about the conclusions that can be drawn from Lambeth's carefully collected data. At most, only 20% of the variation in CLASS or simulation scores represents consistent differences between individual PSTs. At worst, effectively none of the variation in scores represents consistent differences between PSTs. Our results reinforce the challenges raised by Bartanen & Kwok (2021) and suggest they might extend beyond supervisor ratings. Drawing conclusions about individual PSTs' skills based on these scores is risky. When scores stem not from instructional skills but instead from statistical error and/or the systematic conditions under which an observation was conducted (e.g., working with a particular mentor), it would be misguided to use them to make consequential decisions about specific supports for PSTs or an individual's readiness to enter the classroom.

What then are the implications of our results for TPPs? If observational measures of PSTs' teaching provide little information about differences between PSTs, then one possible implication is that we should stop collecting such data. Doing so, however, leaves TPPs to "fly blind" when it comes to understanding how the experiences TPPs provide PSTs develop their instructional skills. Instead, our results highlight several promising strategies that TPPs should consider when implementing such measures.

First, our results suggest the benefit of supplementing global assessments of "instructional quality" with finer-grained measures of finer-grained aspects of teaching. With broad measures like CLASS, differences between PSTs in their use of "feedback loops" (a behavioral indicator in the dimension of "Quality of Feedback") for example, are likely to be drowned out by the myriad skills captured in a domain-score for Instructional Support. Simulation-based measures of fine-grained skills might be better able to detect differences between individual PSTs than CLASS scores. TPPs can also provide more nuanced feedback to

PSTs about finer-grained aspects of teaching like providing timely behavioral redirections than the far more amorphous and multifaceted CLASS dimension of "Positive Climate" let alone the domain score for Emotional Support (Hill & Grossman, 2013; Wylie, 2020).

Second, TPPs can consider strategies for reducing the influence of contextual conditions on PSTs' scores. The higher reliability of Lambeth's simulation scores provides suggestive evidence that raters randomly assigned to observations have little systematic influence on PST scores. This greatly reduces the likelihood that observed differences in scores or growth between PSTs stem from differences in rater standards, a serious issue when PSTs are rated by a single supervisor or mentor (Bartanen & Kwok, 2021).

None of this is to say that supervisors or mentors should not *also* assess and support PSTs, using their knowledge of the context in which a PST is working to inform their understanding of a PST's capacities, contingent on the students with whom they are working. Indeed, mentors and supervisors may be better able to provide formative, context-specific feedback and support for PSTs' skill development if they do not *also* rate PST performance (Papay, 2012). Instead, we argue for a system that would decouple context-specific support systems for PSTs from systems designed to provide precise and actionable insights to TPPs. This approach is costly, however. Lambeth invested in having PSTs video-record their observations and paying additional personnel to score each observation, instead of leveraging supervisor or mentor ratings. Alternatively, TPPs could ask supervisors to collect videos and then randomly assign supervisors to videos. Moreover, we see these costs diminishing dramatically in future years, given rapid advancements to video capture, automated transcription, and even automated scoring of teaching using the textual data culled from such transcripts (Cohen et al., 2023).

Despite up-front investments, minimizing the impact of raters on PSTs' scores could pay dividends in terms of the accuracy of information gleaned from these observations.

TPPs can also standardize the conditions of observations as an alternative to randomization, reducing the likelihood that differences between scores reflect differences between circumstantial conditions. Our results comparing scores from standardized simulations with CLASS scores provides suggestive evidence of the value of this approach. Though simulations are, by design, artificial and do not reflect the full complexity of the classroom, our results suggest that they can be helpful for identifying differences in how candidates enact their knowledge and skills when required to face common problems of practice. Importantly, PSTs complete simulations without the aid of a mentor. While we recognize the importance of observing how PSTs enact their knowledge and skills in real classrooms, we argue that it's equally important for some of the measures we use to capture what PSTs can do on their own, instead of what they do when assisted by mentors.

In practice, clinical observations cannot feasibly be standardized to the same extent as simulations, and so are likely to suffer from more severe reliability and sensitivity issues. Our results indicate these challenges impact multiple CLASS domains. Standardizing would require observing PSTs teach the same lesson within the same mentor's classroom, a logistically infeasible and disruptive approach for PSTs, mentors, and students. Randomization may be logistically more feasible but would limit TPPs' ability to intentionally match PSTs with specific geographic areas, grade levels, content areas, school contexts, or mentor characteristics.[7]

[7] In theory, TPPs could exert some control over these issues by first dividing PSTs into groups based on preferences for geographic area, grade level, content area, school context, and/or mentor characteristics. However, this requires a sufficiently large number of PSTs and mentor classrooms within each grouping to allow for random assignment. The more characteristics a TPP wants to influence, the more groups would be required and the smaller the size of each group. While it may be feasible, therefore, for a TPP to randomly assign PSTs to mentors within geographic areas, it may not be feasible for them to randomly assign PSTs to mentors within content area, grade levels, and geographic areas.

To be clear, our results are about highlighting tradeoffs with these measurement decisions. We do not suggest designing observational systems that produce highly reliable and sensitive data at the expense of PSTs learning in clinical experiences and working alongside skilled mentors (Goldhaber et al., 2022; Ronfeldt, 2012). In addition to reflecting more authentic instructional contexts, observations during clinical placements can use existing TPP systems, whereas simulation-based measures require substantial up-front technology and infrastructure investments. Instead, our argument is that TPPs need to understand the affordances and constraints of the different measures they use, and ideally build a suite of measures with distinct strengths for distinct purposes.

Our results also make clear that mentors matter not just for PSTs' learning (Goldhaber et al., 2022; Ronfeldt et al., 2018), but also for TPPs' understanding of PSTs' development. More work is needed to understand how a mentor's classroom influences PST scores and to develop strategies for TPPs to minimize this influence. Indeed, we are not aware of any other work that highlights this issue or provides potential solutions. At a minimum, TPPs should consider implementing some additional standardization, such as a set of standardized instructional activities to complete at set times during clinical placements (e.g., facilitating a discussion about a word problem or orchestrating an analysis of a historical text). TPPs should also consider exploring new or modified measures whose indicators focus squarely on PST practices to minimize the potential for mentor effects. Measures focused on *classroom* quality that attend to the actions of all adults in the room (like the CLASS) might be less useful in the pre-service context.

Finally, we highlight two potentially helpful strategies for addressing issues of reliability and sensitivity for clinical observations and simulation-based measures. First, our results suggest

estimates of PST growth over time will likely be more reliable than a score at a single point in time, because growth estimates incorporate multiple observation scores. This suggests TPPs would be well-served by collecting multiple scores of PST skill over the course of a program, not just at the end. It also suggests the value of TPPs using measures of growth rather than end-of-program average scores for accreditation and program approval purposes to provide a clearer and more accurate sense of how the TPP supports PST skill development.

Second, statistical adjustments using multi-level models can theoretically correct for low reliability and isolate consistent differences between PSTs. In practice, however, this kind of statistical adjustment requires substantial methodological expertise and represents a non-trivial departure from the status quo. Currently, most TPPs use raw, unadjusted scores for internal data analysis purposes, external reporting to accreditation bodies, and communicating results to PSTs and other stakeholders. Shifting from raw scores to statistically adjusted scores would require TPPs to develop the capacity to conduct such analyses, interpret adjusted scores in their own internal decision-making, and support PSTs and other stakeholders with understanding and interpreting these scores.

TPPs face complex decisions and trade-offs in managing their PST data systems. Systematic collection of PST data requires substantial resources and time investment on the part of TPPs and teacher educators. Altering these systems to address the measurement challenges highlighted here requires even more time and resources, especially when new technologies or more intensive data collection efforts are required. Additionally, TPPs must navigate potential tensions between improving the reliability and sensitivity of PST data and ensuring preparation experiences continue to support PST learning. Standardizing lesson objectives, for example, may

improve reliability, but is also challenging when PSTs work in a wide range of grade levels and school contexts.

While our results raise serious concerns about using observational measures of PSTs' instructional practice to draw conclusions about individual PSTs' instructional skills, they also highlight other kinds of inferences we *can* make, even in the absence of high reliability. The simulation scores, for example, can successfully detect differences between groups of PSTs who had different learning experiences, such as coaching or self-reflection supports. They are sensitive enough to detect these differences immediately after those experiences, as well as months down the road. For programs trying out new courses, clinical experiences, or other kinds of learning supports, such measures could be invaluable in determining the degree to which such programmatic shifts are associated with corresponding shifts in PSTs' skills. CLASS scores are sensitive enough to detect program-wide patterns in PSTs' relative strengths and weaknesses across multiple facets of instruction, in this case, considerably stronger Classroom Organization skills than Instructional Support skills, though we caution both measures also capture a good deal of information about PSTs' mentors, too.

TPPs must decide what data will support the inferences they wish to make. Measures that detect differences between groups of PSTs, for example, may look different than measures that can detect differences between individual PSTs. Obtaining reliable estimates of PST growth over time requires a different observation schedule than obtaining reliable estimates of PSTs' skills at specific moments. Prior work on in-service teacher observations suggests that expecting a single measure to serve several different purposes is unwise (Hill & Grossman, 2013; Papay, 2012). Instead, TPPs should use distinct measures to draw distinct conclusions about PSTs. This means that TPPs must be crystal clear about what conclusions they wish to draw when making

decisions about what measure(s) to use, especially if logistical and financial constraints prevent the use of multiple observational measures.

TPPs stand to reap large benefits when they develop systems that enable data-driven programmatic decision-making. TPPs can better decide how to allocate limited resources to support the PSTs and areas of instructional practice with the greatest need, as well as how to evaluate the effects of specific preparation experiences on PST learning and skill development. However, the details of the data matter if we want TPPs to improve the quality of teacher preparation and not just complete a compliance exercise for program accreditation. This is especially true in the pre-service period where PSTs are likely to exhibit smaller differences in skill than in-service teachers with a range of experience and when contextual factors, such as mentors, may influence assessments of PSTs' skills. To ensure that *all* TPP graduates enter the classroom ready to do the important and complex work of supporting K-12 students, we ultimately need PST data systems that are 1) reliable enough to identify consistent differences between individual PSTs, 2) sensitive enough to detect differences between PSTs and facets of instruction, and 3) allow TPPs to generate empirically-backed conclusions about PST learning and development. Our findings suggest we need considerably more work on all fronts.

References

American Association of Colleges for Teacher Education (AACTE). (2018). A pivot toward clinical practice, its lexicon, and the renewal of educator preparation (Technical Report). http://www.nysed.gov/common/nysed/files/cpc-aactecpcreport.pdf

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review*, *42*(1), 76–98.

Allen, J., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-secondary intervention. *Journal of Research on Educational Effectiveness*, *8*(4), 475–489.

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.

Allen, M., & Coble, C. (2018). Creating a data culture in educator preparation: The role of the states. In E. Mandinach and E. Gummer (Eds.), *Data for continuous programmatic improvement* (pp. 35–67). Routledge.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, *131*(3), 1415–1453.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). "Do first impressions matter? Improvements in early career teacher effectiveness. *AERA Open Access Journal, 1*(4), 1-23.

Bartanen, B., & Kwok, A. (2020). *Pre-service teacher quality and workforce entry.* EdWorkingPaper No. 20-223. Annenberg Institute at Brown University.

Bartanen, B., & Kwok, A. (2021). Examining clinical teaching observation scores as a measure of preservice teacher quality. *American Educational Research Journal*, *58*(5), 887–920.

Bassok, D., Magouirk, P., & Markowitz, A. J. (2021). Systemwide Quality Improvement in Early Childhood Education: Evidence From Louisiana. *AERA Open*, *7*(1), 1-19.

Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education*, 59, 1–12.

Bastian, K. C., Lys, D., & Pan, Y. (2018). A framework for improvement: analyzing performance-assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education*, *69*(5), 448–462.

Bastian, K. C., Lys, D. B., & Whisenant, W. R. L. (2023). Does placement predict performance? Associations between student teaching environments and candidates' performance assessment scores. *Journal of Teacher Education*, *74*(1), 40-54.

Bastian, K. C., Patterson, K. M., & Carpenter, D. (2022). Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy*, *36*(7), 1583-1611.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, *30*(1), 3–29.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). John Wiley & Sons.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, *27*(4), 793–818.

Briggs, D. C., & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement*, *79*(4), 636–664.

Campbell, S.L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233-1267.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, *75*(2), 311–337.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5–51.

Charney, R. S. (1993). *Teaching children to care: Management in the responsive classroom*. Northeast Foundation for Children.

Cohen, J., Erickson, S., Krishnamachari, A., & Wong, V. (2023). Experimental evidence on the robustness of coaching supports in teacher education. *Educational Researcher*.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378–387.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, 43(2), 208–231.

Council for the Accreditation of Educator Preparation (CAEP). (2022). 2022 CAEP Standards.

Davis, S. C., & Peck, C. A. (2020). Using data for program improvement in teacher education: A study of promising practices. *Teachers College Record*, *122*(3), 1–48.

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.

Feuer, M., Floden, R., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. National Academy of Education.

Forzani, F. M. (2014). Understanding "core practices" and "practice-based" teacher education: Learning from the past. *Journal of Teacher Education*, *65*(4), 357–368.

Gitomer, D. (Ed.). (2009). *Measurement issues and assessment for teaching quality*. Sage Publications.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1–32.

Goldhaber, D. (2019). Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education*, *70*(2), 90–101.

Goldhaber, D., Krieg, J., & Theobald, R. (2020). Effective like me? Does having a more productive mentor improve the productivity of mentees? *Labour Economics, 63*, 101792

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.

Goldhaber, D., Ronfeldt, M., Cowan, J., Gratz, T., Bardelli, E., & Truwit, M. (2022). Room for improvement? Mentor teachers and the evolution of teacher preservice clinical evaluations. *American Educational Research Journal*, *59*(5), 1011-1048.

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

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system-secondary. *The Journal of Early Adolescence*, *35*(5–6), 651–680.

Henry, G. T., Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., Lys, D. B., & Covington, V. M. (2013). The predictive validity of measures of teacher candidate programs and performance: Toward an evidence-based approach to teacher preparation. *Journal of Teacher Education*, *64*(5), 439–453.

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, *83*(2), 371–384.

Hill, H., Mancenido, Z., & Loeb, S. (2020). *New research for teacher education.* EdWorkingPaper No. 20-252. Annenberg Institute at Brown University.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Bill & Melinda Gates Foundation.

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., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234–249.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal*, *104*(5), 409–426.

Lampert, M., & Graziani, F. (2009). Instructional activities as a tool for teachers' and teacher educators' learning. *The Elementary School Journal*, *109*(5), 491–509.

Levine, A. (2006). Educating school teachers *(No. 2).* Washington, DC: The Education Schools Project. http://www.edschools.org/teacher_report_release.htm.

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, *102*(4), 916–932.

Mancenido, Z. (2022). *Impact evaluations of teacher preparation practices: Challenges and opportunities for more rigorous research.* EdWorkingPaper No. 22–534. Annenberg Institute at Brown University.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378–386.

Pacheo, A. (2009). Mapping the terrain of teacher quality. In D. Gitomer (Ed.), *measurement issues and assessment for teaching quality* (pp. 168–178). Sage Publications.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review*, *82*(1), 123–141.

Paris, D., & Alim, H. S. (2014). What are we seeking to sustain through culturally sustaining pedagogy? A loving critique forward. *Harvard Educational Review, 84*(1), 85–100. https://doi.org/10.17763/haer.84.1.982l873k2ht16m77

Park, Y. S., Chen, J., & Holtzman, S. L. (2015). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 381-414). John Wiley & Sons.

Peck, C. A., & McDonald, M. (2013). Creating "cultures of evidence" in teacher education: Context, policy, and practice in three high-data-use programs. *The New Educator*, *9*(1), 12–28.

Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, 121, 33–46.

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.

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis*, *34*(1), 3–26.

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving preservice teachers' feelings of preparedness to teach through recruitment of instructionally effective and experienced cooperating teachers: A randomized experiment. *Educational Evaluation and Policy Analysis*, *42*(4), 551–575.

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*(7), 405–418.

Ronfeldt, M., Matsko, K. K., Greene Nolan, H., & Reininger, M. (2021). Three different measures of graduates' instructional readiness and the features of preservice preparation that predict them. *Journal of Teacher Education*, *72*(1), 56–71.

Texas State Legislature. Texas Senate Bill 174 (2009).

US Department of Education. (2011). Our future, our teachers: The Obama administration's plans for teacher education reform and improvement. Washington, D.C.

van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50, 88–95.

Wallace, T. L., Parr, A. K., & Correnti, R. J. (2020). Assessing teachers' classroom management competency: A case study of the classroom assessment scoring system–secondary. *Journal of Psychoeducational Assessment*, *38*(4), 475–492.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project.

Wylie, E. C. (2020). Observing formative assessment practice: Learning lessons through validation. *Educational Assessment*, *25*(4), 251–258.

Tables

Table 1: Proportion of the variation that reflects consistent differences between PSTs relative to the overall variation in scores.[8]

	CLASS: Instructional Support	**CLASS:** Classroom Organization	**SIM:** Redirection	**SIM:** Text-Focused Instruction
Overall variation, assuming no growth in scores over time	0.07	0.03	0.17	0.11
Overall variation, allowing for growth in scores over time	0.04	0.03	0.22	0.21
Variation in baseline scores	0.04			0.21
Variation in growth over time	0.15			0.15

Note: Estimates are calculated using Equations 1-4 described in Appendix B. Estimates for fixed effects, random effects, and variance components from these equations can be found in Tables C1-C3.

Table 2. Proportion of score variation that reflects consistent differences between PSTs and proportion of score variation that reflects consistent differences between mentors (for CLASS scores) and raters (for simulation scores).

	CLASS Scores		Simulation-based Scores	
	Instructional Support	Classroom Organization	Redirection	Text-Focused Instruction
PST baseline	0.0002	0.00	0.23	0.21
PST growth	0.0004	n/a	n/a	0.15
Rater/Mentor	0.09	n/a	0.03	0.01
Sources of variation	Mentors PSTs Residual error	Mentors PSTs Residual error	Raters PSTs Residual error	Raters PSTs Residual error

Note: Estimates are calculated following the methodology described in Appendix B under RQ2&3.

[8] For Classroom Organization and Redirection scores all PSTs effectively grow at the same rate, so we cannot estimate what proportion of this "growth" variation reflects consistent differences between-PSTs.

Table 3. Spearman's rank correlations comparing predicted CLASS and simulation scores across the different facets of instruction.

	CLASS Scores		Simulation Scores	
	Without controls for mentor	With controls for mentor	Without controls for rater	With controls for rater
Observation 0	0.45***	0.45***	0.25*	0.25*
Observation 1	0.44***	0.41***	0.24*	0.24*
Observation 2	0.43***	0.41***	-0.14	-0.14
Observation 3	0.42***	0.41***	-0.01	-0.01

*** p<0.01, ** p<0.05, * p<0.1. Note: CLASS Observations 0-4 are scored on *both* Classroom Organization and Instructional Support. Sim Observations 0-1 are scored on Redirection and 2-3 on Text-Focused Instruction.

Figures

Figure 1. Conceptual framework articulating how data on PST knowledge, skills, and dispositions can contribute to improvements in teacher preparation.

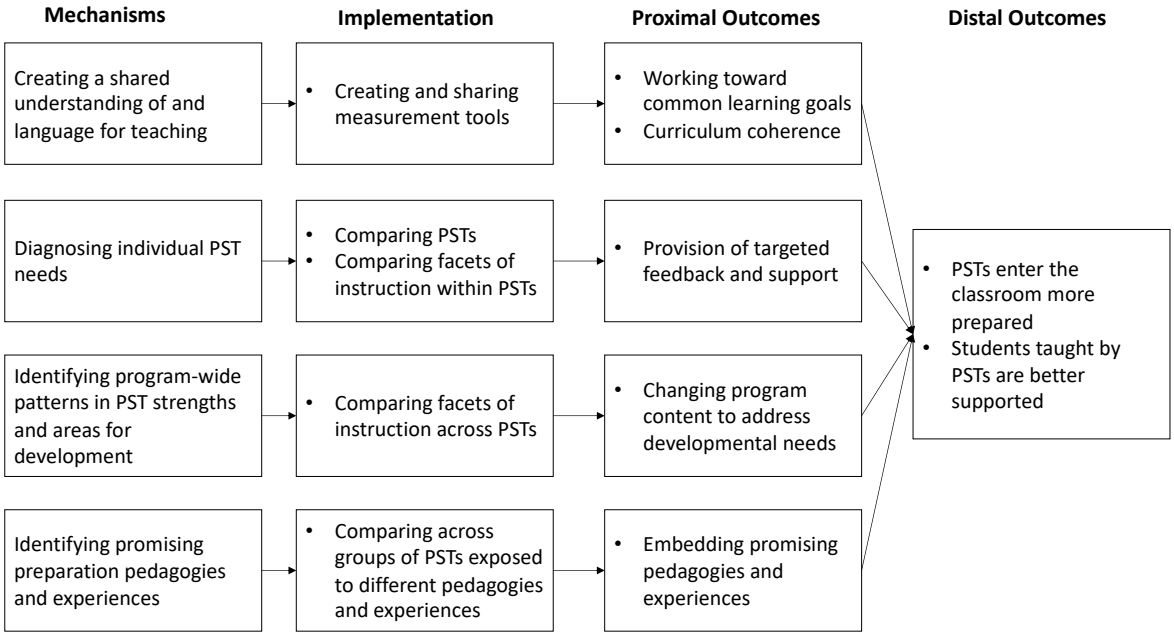
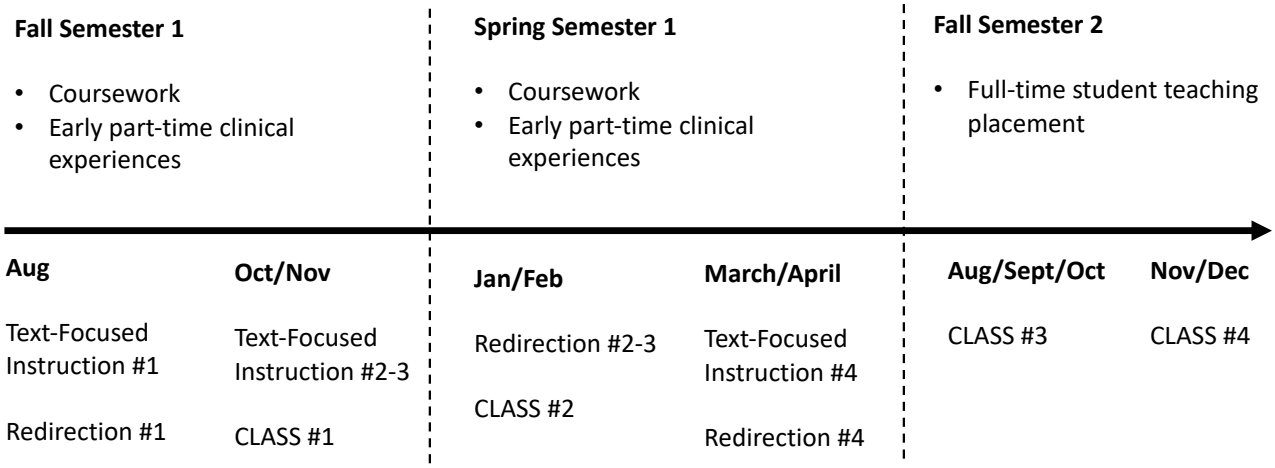


Figure 2. Data collection timeline for each simulation-based measure and CLASS scores.



Note: this timeline reflects the collection of CLASS scores for the cohort that entered Lambeth in 2017-18 and the collection of simulation scores for the cohort that entered Lambeth in 2018-19 to be most reflective of our analytic sample. While the data collection timelines for other years were similar, there were some differences. For example, only the first two CLASS scores were collected in 2018-19 and 2019-20 because of logistical constraints and the Covid-19 pandemic, respectively. Similarly, 2017-18 was a pilot year for the simulation-based measures that did not include the initial baseline data collection in August. Additionally, the fourth simulation-based observations in March/April did not occur in 2019-20 because of the Covid-19 pandemic. The timeline for coursework and clinical experiences, however, was similar across all three years, except for a small subset of PSTs who entered Lambeth in 2018-19 or 2019-20 and completed a one-year MAT in Secondary Education. These students began coursework in the summer before the first Fall Semester, completed coursework and early part-time clinical experiences in the first Fall Semester, and completed a full-time student teaching placement in the first Spring semester. For these students, the first CLASS observation was collected from their early clinical experience and the second was collected during their full-time student teaching placement. For other students who entered in 2018-19 and 2019-20 all available CLASS observations were collected during their early part-time clinical experiences.

Figure 3. Predicted PST scores after accounting for measurement error to isolate between-PST variation in scores.

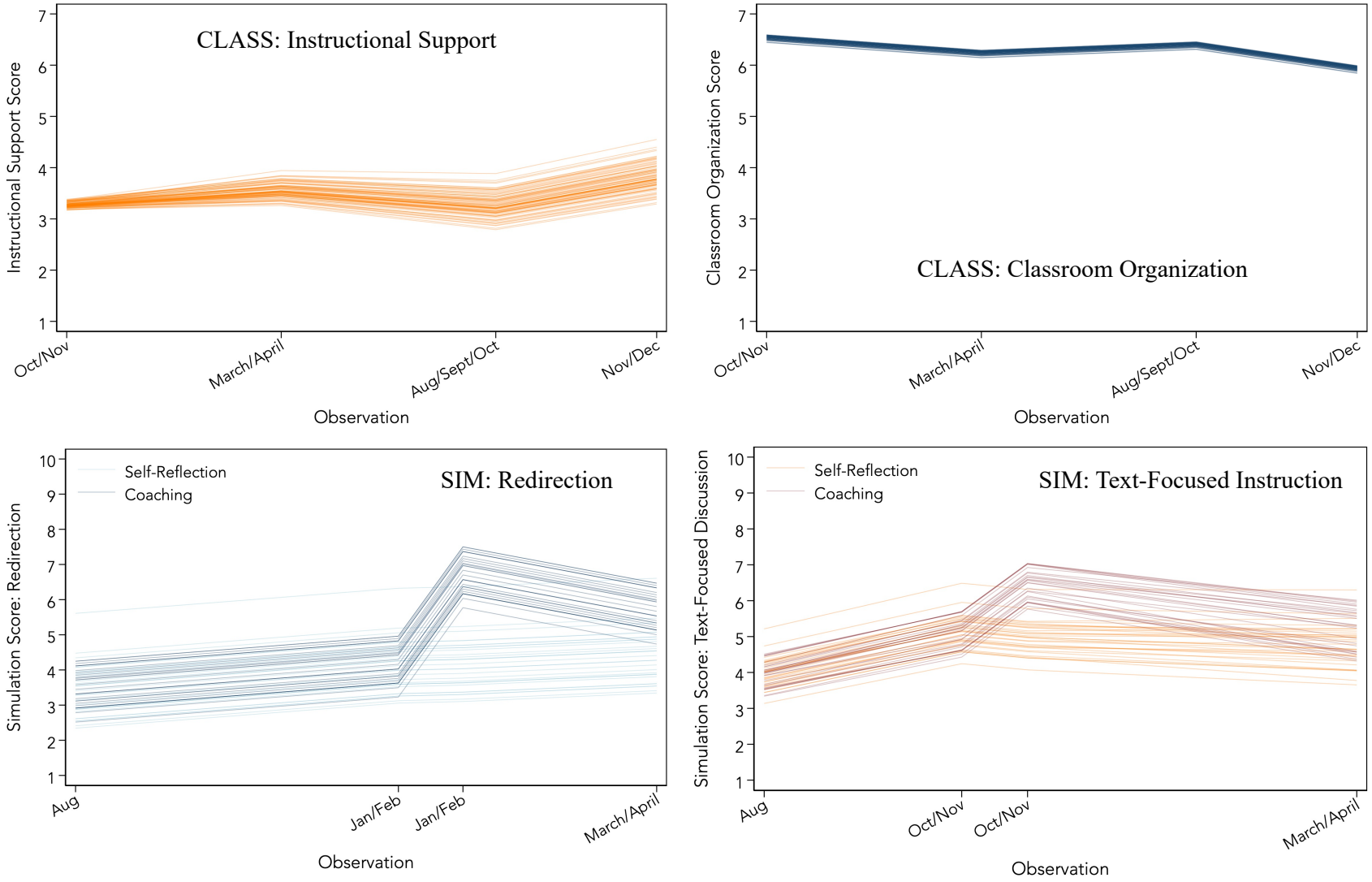


Figure 4. Comparison between Instructional Support and Classroom Organization scores using predicted scores.

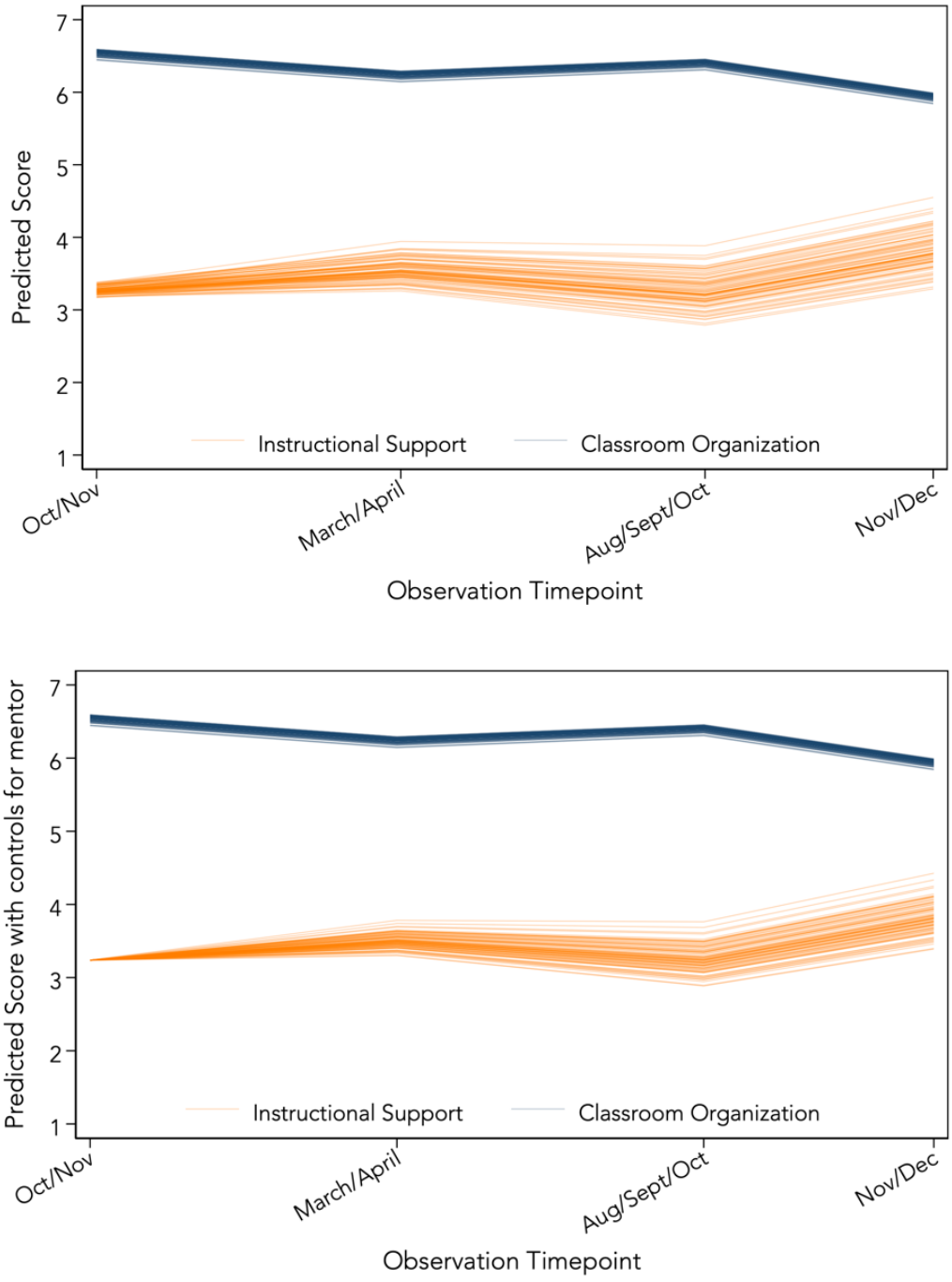
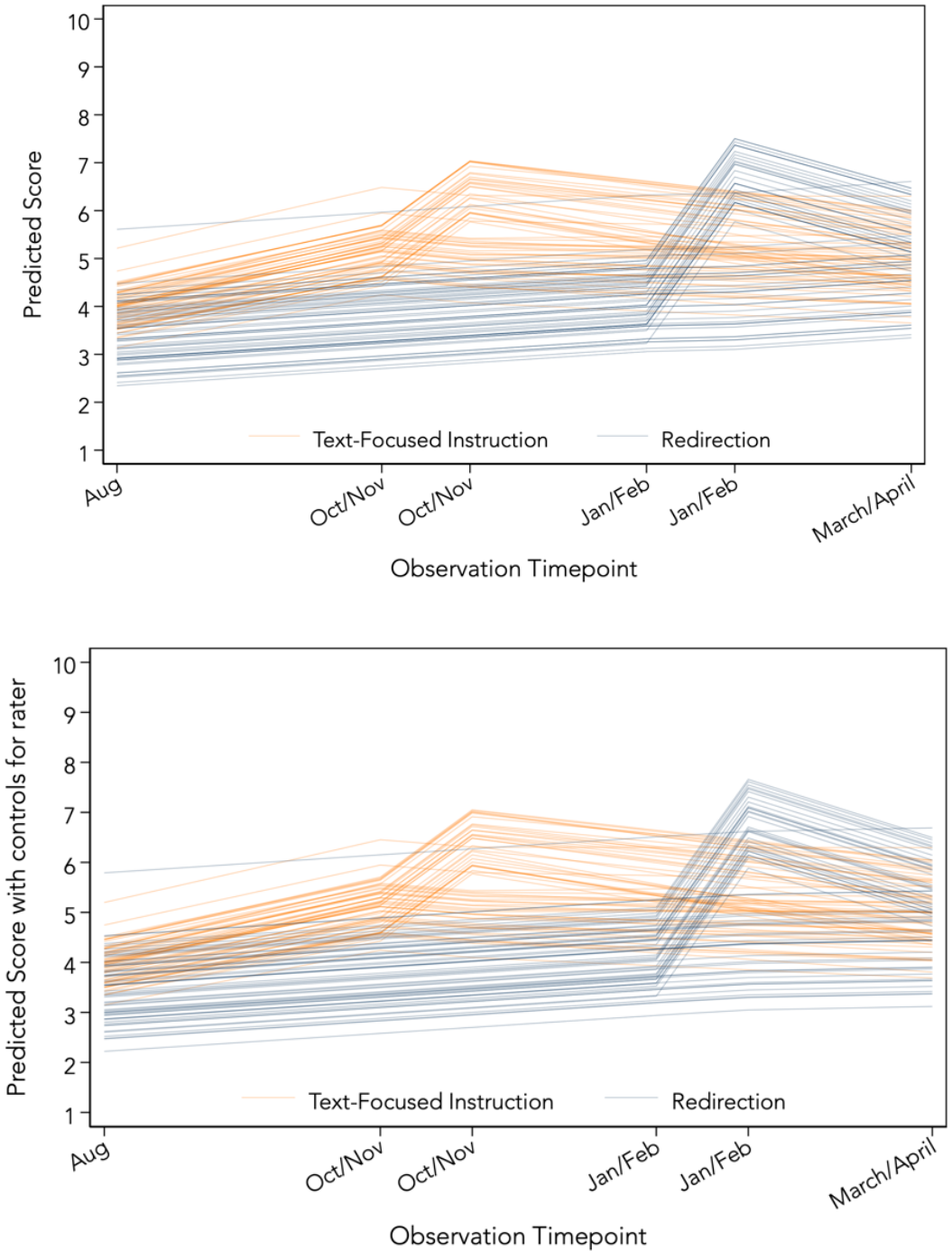


Figure 5. Comparison between Text-Focused Instruction and Redirection scores using predicted scores. Both models include controls for coaching treatment status.



Appendices

Appendix A: Selecting our Analytic Sample

Our full data set includes both CLASS and simulation-based scores from three academic years. Complete data for a single PST consists of four CLASS observations (including both Classroom Organization and Instructional Support), four Redirection observations, and four Text-Focused Instruction observations, for a total of 12 distinct observation timepoints.[9] Table A1 illustrates the CLASS and simulation data we have available in our full dataset, highlighting the number of PSTs with complete data and the extent to which there is missing data. To allow comparisons across measures and avoid bias from missing data, our analytic sample would ideally consist of all PSTs with complete data. Unfortunately, there are no PSTs in any year for whom we have complete CLASS and simulation data because of changes in the number of observations conducted across years and the disruptions to data collection caused by the Covid-19 pandemic in spring 2020. It is also not possible to expand our analysis to PSTs with fewer scores, because four scores reflect a minimum requirement to have adequate coverage across PSTs' program (i.e., covers a reasonable part of the time the PSTs are in the program rather than a small slice) and to conduct the kind of multi-level models we use to answer our first two research questions and analyze changes in PSTs' instructional skills over time (Raudenbush & Bryk, 2002). Instead, we are left with 83 PSTs for whom we have complete CLASS data in 2017-18 and 60 PSTs for whom we have complete simulation data in 2018-19 as the best approximation of the ideal analytic sample. We acknowledge that these PSTs may not be

[9] For each of the two CLASS domains, observation scores were generated from the same video at the same timepoint, resulting in a total of 8 scores but only 4 observations. There are no PSTs for whom different numbers of Classroom Organization and Instructional Support scores are available. For each of the two simulation-based scores, however, PSTs were observed at different times using different videos, thus it is possible for a PST to have complete Text-Focused Instruction data (4 scores), while also having incomplete Redirection data (less than 4 scores).

representative of the broader population of PSTs enrolled at Lambeth between 2016-17 and 2019-20, though the data we have does not suggest they are, in any way, different from the broader population. However, given our focus on the measurement properties of CLASS and simulation scores, where more observation timepoints can only improve measurement quality and fewer observation timepoints can only decrease measurement quality, we feel confident that the results from our analytic sample provide a reasonable estimate of the best-case scenario in terms of reliability and sensitivity for CLASS and simulation scores.

Table A1. Missing CLASS and simulation scores for our full dataset across three years.

Year	Number of PSTs with complete data (12 observations)	Number of PSTs with complete CLASS data, but incomplete simulation data	Number of PSTs with complete simulation data, but incomplete CLASS data	Number of PSTs with incomplete simulation and incomplete CLASS data	Total
2017-18	0	83[10]	0	29	118
2018-19	0	0	60	49	109
2019-20	0	0	0	107	107

Since the analytic sample used for our analysis of CLASS scores differs from the analytic sample used for our analysis of simulation scores, we also explore in Table A3 the extent to which PSTs' demographic characteristics differ across these two samples. In doing so, we find few significant differences between PSTs, and the p-value for the joint F-test indicates that we cannot reject the null hypothesis that all differences in demographics by sample are equal to zero. Nonetheless, we do observe that PSTs in the simulation sample from 2018-19 are more likely to

[10] There are 6 additional PSTs for whom all four CLASS observations are available. However, the date of observation tied to these videos does not match the timeline of observation set by Lambeth. After discussing these PSTs with Lambeth staff, we decided to exclude these PSTs as outliers.

be female, less likely to be earning their credential in Elementary education (as opposed to Secondary or Special Education) and more likely to be under the age of 27.

Table A2. Covariate balance table estimating the differences in demographic characteristics between PSTs included in our analytic sample and the broader population.

	CLASS			Simulation			CLASS and Simulation		
	Analytic sample mean	Difference	N	Analytic Sample Mean	Difference	N	Analytic Sample Mean	Difference	N
High School GPA	3.407***	0.0521	70	3.460***	0.0273	94	3.532***	-0.0734	796
	(0.0637)	(0.0780)		(0.0552)	(0.0711)		(0.0123)	(0.0460)	
Enrolled in Master's	0.543***	-0.157	11	0.592***	-0.159	109	0.405***	-0.0199	1,067
degree program	(0.0849)	(0.101)	8	(0.0709)	(0.0959)		(0.0157)	(0.0557)	
Enrolled in Secondary	0.657***	-0.27***	11	0.551***	-0.30***	109	0.435***	-0.0494	1,067
Education track	(0.0809)	(0.0972)	8	(0.0717)	(0.0913)		(0.0158)	(0.0558)	
Enrolled in Special	0.114**	0.00620	11	0.163***	-0.0966	109	0.151***	-0.0309	1,067
Education track	(0.0542)	(0.0651)	8	(0.0533)	(0.0624)		(0.0114)	(0.0375)	
Home language is	0.895***	0.00911	71	0.947***	-0.0526	95	0.894***	0.00984	269
English	(0.0714)	(0.0826)		(0.0366)	(0.0550)		(0.0210)	(0.0461)	
Female	0.429***	0.138	11	0.714***	0.119	109	0.450***	0.116**	1,067
	(0.0844)	(0.101)	8	(0.0651)	(0.0812)		(0.0159)	(0.0567)	
White	0.857***	-0.00172	11	0.857***	-0.0238	109	0.850***	0.00583	1,067
	(0.0597)	(0.0712)	8	(0.0505)	(0.0700)		(0.0114)	(0.0403)	
Under 27 years old	0.514***	0.0881	11	0.735***	0.115	109	0.583***	0.0191	1,067
	(0.0852)	(0.101)	8	(0.0637)	(0.0788)		(0.0157)	(0.0560)	
Attended HS where	0.474***	0.0263	71	0.605***	-0.149	95	0.521***	-0.0207	269
students were	(0.116)	(0.136)		(0.0801)	(0.104)		(0.0340)	(0.0775)	
majority white									
Attended HS where	0.316***	-0.00206	70	0.184***	0.0614	95	0.281***	0.0326	268
students had high	(0.108)	(0.127)		(0.0636)	(0.0858)		(0.0306)	(0.0721)	
socio-economic status									
Attended a public HS	0.842***	-0.0152	71	0.842***	-0.0273	92	0.829***	-0.00246	263
	(0.0849)	(0.100)		(0.0598)	(0.0802)		(0.0260)	(0.0587)	
Attended HS where	0.579***	-0.0597	71	0.342***	0.158	94	0.500***	0.0192	268
students received high	(0.115)	(0.135)		(0.0778)	(0.103)		(0.0341)	(0.0775)	

Different methods for assessing pre-service teachers' instruction: Why measures matter

	CLASS			Simulation			CLASS and Simulation		
	Analytic sample mean	Difference	N	Analytic Sample Mean	Difference	N	Analytic Sample Mean	Difference	N
scores on achievement tests									
Plans to teach in a school where students are majority white	0.0526	-0.0142	71	0.0526	0.000940	94	0.056***	-0.0171	268
	(0.0520)	(0.0586)		(0.0366)	(0.0476)		(0.0156)	(0.0310)	
Plans to teach in a school where students receive high scores on achievement tests	0.579***	-0.194	71	0.289***	0.0677	94	0.39***	-0.00890	268
	(0.115)	(0.134)		(0.0744)	(0.0986)		(0.0334)	(0.0755)	
Plans to teach in a school where students have high socio-economic status	0.0526	0.0243	71	0.0526	-0.0169	94	0.051***	0.0260	268
	(0.0520)	(0.0641)		(0.0366)	(0.0444)		(0.0150)	(0.0400)	
	Joint F-test	p = 0.11		Joint F-test	p = 0.20		Joint F-test	p = 0.11	796

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A3. Covariate balance table estimating the differences in demographic characteristics between PSTs in the CLASS analytic sample and PSTs in the simulation analytic sample.

	Maximum Sample			Consistent Sample (N=101)	
	CLASS sample mean	Difference	N	CLASS sample mean	Difference
High School GPA	3.459***	0.0285	107	3.453***	0.0220
	(0.0447)	(0.0633)		(0.0460)	(0.0648)
Enrolled in Master's degree program	0.386***	0.0478	143	0.340***	0.0914
	(0.0538)	(0.0839)		(0.0692)	(0.0974)
Enrolled in Secondary Education track	0.386***	-0.136*	143	0.420***	-0.165*
	(0.0538)	(0.0779)		(0.0665)	(0.0935)
Enrolled in Special Education track	0.120***	-0.0538	143	0.00	0.0784**
	(0.0360)	(0.0484)		(0.0273)	(0.0384)
Home language is English	0.904***	-0.00911	109	0.900***	-0.0176
	(0.0413)	(0.0582)		(0.0445)	(0.0626)
Female	0.566***	0.267***	143	0.900***	-0.0176
	(0.0548)	(0.0731)		(0.0445)	(0.0626)
White	0.855***	-0.0221	143	0.760***	0.0439
	(0.0389)	(0.0621)		(0.0589)	(0.0829)
Under 27 years old	0.602***	0.248***	143	0.960***	-0.0384
	(0.0541)	(0.0713)		(0.0337)	(0.0474)
Attended HS where students were majority white	0.500***	-0.0439	109	0.520***	-0.0886
	(0.0700)	(0.0966)		(0.0711)	(0.1000)
Attended HS where students had high socio-economic status	0.314***	-0.0681	108	0.320***	-0.0455
	(0.0656)	(0.0873)		(0.0652)	(0.0917)
Attended a public HS	0.827***	-0.0121	106	0.820***	0.00353
	(0.0530)	(0.0752)		(0.0547)	(0.0769)
Attended HS where students received high scores on achievement tests	0.519***	-0.0192	108	0.540***	-0.0302
	(0.0699)	(0.0972)		(0.0713)	(0.100)
Plans to teach in a school where students are majority white	0.0385	0.0151	108	0.0400	0.0188
	(0.0269)	(0.0406)		(0.0310)	(0.0436)

	Maximum Sample			Consistent Sample (N=101)	
	CLASS sample mean	Difference	N	CLASS sample mean	Difference
Plans to teach in a school where students receive high scores on achievement tests	0.385***	-0.0275	108	0.380***	-0.0663
	(0.0681)	(0.0939)		(0.0678)	(0.0954)
Plans to teach in a school where students have high socio-economic status	0.0769**	-0.0412	108	0.0800**	
	(0.0373)	(0.0449)		(0.0336)	
					-0.0408
					(0.0473)
				Joint F-test	p = 0.526

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Appendix B: Methodological Appendix

RQ1: To what extent does each measure provide the intended information about differences in PSTs' instructional practice?

To answer this question, we fit a multi-level model to model PST scores as a function of the number of months since the first observation and then partition the variance in CLASS and simulation scores. Our primary specification is the two-level model shown in Equations 1 and 2.

$$(1) \qquad y_{tj} = \pi_{0j} + \pi_{1j} + \pi_{2j} + \pi_{3j} + \varepsilon_{tj}$$

$$(2) \qquad \pi_{0j} = \beta_{00} + u_{0j}$$

$$\pi_{tj} = \beta_{tj} + u_{1j}$$

In the Level 1 model (Equation 1), y_{tj} represents PST *j*'s CLASS or simulation score at time *t*, and is modelled as a function of a PST's baseline score at t=0 (π_{0j}), three time splines, centered at the first observation (t=0) and capturing the number of months since the first observation, to flexibly allow for non-linear changes in scores at each subsequent observation timepoint (π_{1j}, π_{2j}, π_{3j}) and a residual error term, (ε_{tj}), where $\varepsilon_{tj} \sim N(0, \sigma^2)$. Level 2 models between-PST variation in the Level 1 parameters as shown in Equation 2. Specifically, we estimate each individual PSTs' baseline score (intercept) as a function of the mean baseline score (β_{00}) and a random effect representing each PSTs' unique deviation from the mean (u_{0j}). We also estimate each PSTs' trajectory over time (slope) as a function of the mean slope at time *t* (β_{tj}) and a random effect representing each PSTs' unique time-invariant deviation from the mean trajectory (u_{1j}), where empirically supported.[11] Additionally, for models with simulation outcomes we

[11] We do not include a time random effect for analyses of simulation scores for the redirection scenario or CLASS Classroom Organization scores because empirical results and model fit

include interactions between an indicator for whether the PST received coaching and any post-coaching timepoints to allow PSTs that received coaching to exhibit different trajectories in line with prior work demonstrating coaching treatment effects (Author., 2020; Author, 2021).

We then use the estimated variance of each random effect to calculate the proportion of the variation that reflects consistent differences between PSTs, using two methods. First, we calculate reliability as the proportion of the total variance explained by PST random effects and time random effects as shown in Equation 3, following traditional definitions of reliability (Raudenbush & Bryk, 2002). Second, we follow Briggs & Alzen's (2019) recently proposed approach for calculating the reliability of PST growth, which requires estimating a separate proportion for initial differences between PSTs (intercept), shown in Equation 4, and differences in PST growth over time (slope), shown in Equation 5.

$$(3)\ \rho = 100 * \frac{\sigma^2_{u0j} + \sigma^2_{uj}}{\sigma^2_{u0j} + \sigma^2_{uj} + \sigma^2_{\varepsilon}}$$

$$(4)\ \rho(\pi_{0j}) = \frac{\sigma^2_{u0j}}{\sigma^2_{u0j} + \sigma^2_{\varepsilon}}$$

$$(5)\ \rho(\pi_{tj}) = \frac{\sigma^2_{uj}}{\sigma^2_{uj} + {\sigma^2_{\varepsilon}}/{SST}}$$

We report these as proportions rather than percentages to reflect the fact that they cannot be added together to total 100% of the variation since these equations incorporate different denominators. In Equation 5, SST is calculated as $\sum_{t=0}^{T}(Time_{tj} - \overline{Time_j})^2$ and serves to adjust the proportion to account for the number and spacing of observations. Holding the time-period over which growth is estimated constant, conducting more observations and/or ensuring that

statistics suggest that there is not sufficient variation to warrant the inclusion of a time random effect.

observations are more widely spread out will increase the SST, and therefore increase the estimated proportion. The intuition here is that scores from more or more widely spread-out observations will be more representative of each PST's growth over time than scores from fewer or more narrowly spaced observations. Since the exact spacing between observations varies across PSTs, we calculate the SST using the modal number of months between observations, resulting in an SST of 101 for Instructional Support and 26.75 for the Text-Focused Instruction simulation. We do not estimate Equation 5 for Classroom Organization and the Redirection simulation scores because we cannot estimate time random effects for these models.

RQ2 &3: To what extent do raters and mentors influence PST scores?

To answer this question, we modify Equations 1 and 2 to further decompose the measurement error identified into distinct sources. Specifically, we include a rater random effect for simulation scores to capture systematic patterns in raters' influence on scores and a mentor teacher random effect for CLASS scores to capture systematic patterns in how the placement context influences scores. Because PSTs are not perfectly nested within raters or mentors, we add these random effects as crossed effects, where possible. However, for Classroom Organization and Redirection scores, the crossed model does not converge. As an alternative, we therefore estimate a simplified version of the model with raters or mentors as fixed rather than random effects. We then recalculate Equations 3 and 4, adding the variance of the rater or mentor random effect to the denominator where necessary. We also use these equations to calculate the proportion of the variation explained by rater or mentor effects.

RQ4: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

To explore the sensitivity of CLASS and simulation scores to individual differences, we review results from RQ1with the aim of understanding the extent to which we can detect between-PST variation in scores and trajectories. Specifically, we use log likelihood tests to compare models with and without PST intercept and slope random effects to determine whether there is significant between-PST variation in baseline scores and growth trajectories. We also generate graphs of PST trajectories, using both raw scores and the predicted Best Linear Unbiased Predictors (BLUPs) from RQ1 models to allow visual inspection of how much scores and trajectories vary across PSTs. We use BLUP estimates from RQ1 models without controls for rater or mentor effects as an upper bound of sensitivity.

To explore the sensitivity of scores to group differences, we review results from RQ1 for simulation scores, where a control for participation in coaching vs. self-reflection allows us to evaluate whether simulation scores are sufficiently sensitive to detect differences between PSTs exposed to these different preparation experiences. Unfortunately, we are not aware of any systematic differences in preparation experiences that can be used to compare groups for the CLASS scores.

To explore the sensitivity of scores to differences between different facets of instructional skill, we visually compare the magnitude of Classroom Organization and Instructional Support scores, using estimated BLUPs. We also estimate the Spearman's rank correlation between Classroom Organization and Instructional Support scores at each timepoint. We then apply the same methods to the simulation scores to explore the relationship between Redirection and Text-Focused Instruction scores.

Appendix C: Complete results from multi-level models estimated in RQ1 and RQ1.

Table C1. Coefficients from multi-level models for Instructional Support (Equations 1-2).

	(1)		(2)		(3)	
	Instructional Support					
Fixed Effects						
Time 0	3.27*	(0.08)	3.23*	(0.09)	3.24*	(0.08)
Time 1	0.28*	(0.11)	0.27*	(0.11)	0.27*	(0.11)
Time 2	-0.30*	(0.11)	-0.24*	(0.12)	-0.25*	(0.12)
Time 3	0.57*	(0.11)	0.57*	(0.11)	0.57*	(0.11)
Random Effects (S.D.)						
PST intercept	0.15				0.02	
PST slope	0.03				0.03	
Residual	0.71		0.71		0.69	
Mentor Teacher			0.30		0.22	
Variance Components						
Overall Reliability	0.04		n/a		0.0002	
Intercept Reliability	0.04		n/a		0.0004	
Slope Reliability	0.15		n/a		0.14	
Mentor Teacher	n/a		0.15		0.09	
Residual	0.96		0.85		0.91	
Observations (PSTs)	332	(83)	332	(83)	332	(83)

Robust Standard errors in parentheses, * p<0.05

Table C2. Coefficients from multi-level models for Classroom Organization (Equations 1-2).

	(1)		(2)		(3)	
	Classroom Organization					
Fixed Effects						
Time 0	6.54*	(0.06)	6.54*	(0.07)	6.48*	(0.10)
Time 1	-0.30*	(0.09)	-0.28*	(0.08)	-0.25*	(0.09)
Time 2	0.16	(0.09)	0.15	(0.09)	0.04	(0.18)
Time 3	-0.47*	(0.09)	-0.47*	(0.08)	-0.47*	(0.08)
Mentor fixed effects					x	
Random Effects (S.D.)						
PST intercept	0.10					
Residual	0.55		0.52		0.00	
Mentor Teacher			0.23		0.51	
Variance Components						
Intercept Reliability	0.03		n/a		0.00	
Mentor Teacher	n/a		0.17		n/a	
Residual	0.97		0.83		1.00	
Observations (PSTs)	332	(83)	332	(83)	332	(83)

Robust Standard errors in parentheses * p<0.05

Table C3. Coefficients from multi-level models for simulation-based scores. (Equations 1-2).

	(1)		(2)		(3)		(4)	
	Text-Focused Instruction Simulation				Redirection Simulation			
Fixed Effects								
Time 0	3.97*	(0.16)	3.95*	(0.17)	3.44*	(0.24)	3.47*	(0.28)
Time 1	1.17*	(0.20)	1.16*	(0.20)	0.71*	(0.29)	0.72*	(0.29)
Time 2	-0.17	(0.24)	-0.15	(0.24)	0.04	(0.37)	0.11	(0.36)
Time 3	-0.25	(0.28)	-0.24	(0.29)	0.24	(0.41)	0.07	(0.43)
Time 2 * Coaching	1.51*	(0.30)	1.49*	(0.30)	2.50*	(0.45)	2.45*	(0.44)
Time 3 * Coaching	-0.98*	(0.41)	-0.96*	(0.41)	-1.28*	(0.59)	-1.23*	(0.58)
Random Effects (S.D.)								
PST intercept	0.56		0.56		0.86		0.89	
PST slope	0.09		0.09					
Residual	1.08		1.07		1.61		1.57	
Rater			0.14				0.33	
Variance Components								
Overall Reliability	0.21		0.21					
Intercept Reliability	0.21		0.21		0.22		0.23	
Slope Reliability	0.15		0.15					
Rater	n/a		0.01		n/a		0.03	
Residual	0.79		0.78		0.78		0.73	
Observations (PSTs)	240	(60)	240	(60)	240	(60)	240	(60)

Robust Standard errors in parentheses, * $p<0.05$

Appendix D. Supplementary analyses of sensitivity to between-PST differences in score and growth

Table D1. Between-PST variation in baseline CLASS scores and growth trajectories, accounting for measurement error.

	16th percentile	50th percentile	84th percentile	16th percentile	50th percentile	84th percentile
	CLASS: Instructional Support			CLASS: Classroom Organization		
Estimated trajectories						
Baseline	3.12	3.27	3.42	6.44	6.54	6.64
+ 6 months	0.25	0.28	0.31	-0.30		
+ 5 months	-0.33	-0.30	-0.27	0.16		
+ 2 months	0.54	0.57	0.60	-0.47		
Statistical Significance						
Baseline differences (PST intercept random effects)		0.0324			0.4861	
Differences in growth (PST slope random effects)		0.0056			n/a	

Note: We estimate these exemplar PST trajectories using the random effects coefficients included in Tables B1-B2 and estimated using Equations 1-2 as described in the methodological appendix (Appendix A). P-values are calculated using likelihood ratio tests to evaluate the significance of between-PST differences as described in Appendix A, under RQ3.

Table D2. Between-PST variation in baseline simulation scores and growth trajectories, accounting for measurement error.

	16th percentile	50th percentile	84th percentile	16th percentile	50th percentile	84th percentile
	Simulation: Text-Focused Instruction			Simulation: Redirection		
Estimated trajectories						
Baseline	3.41	3.97	4.52	2.59	3.44	4.30
+ 6 months	1.09	1.17	1.26		0.71	
+ 5 months	-0.26 SR 1.25 C	-0.17 SR 1.34 C	-0.08 SR 1.42 C		0.04 SR 2.54 C	
+ 2 months	-0.34 SR -1.32 C	-0.25 SR -1.23 C	-0.16 SR -1.15 C		0.24 SR -1.03 C	
Statistical Significance						
Baseline differences (PST intercept random effects)		0.0002			0.0002	
Differences in growth (PST slope random effects)		0.5302			n/a	

Note: We estimate these exemplar PST trajectories using the random effects coefficients included in Table B3 and estimated using Equations 1-2 as described in the methodological appendix (Appendix A). P-values are calculated using likelihood ratio tests to evaluate the significance of between-PST differences as described in Appendix A, under RQ3. Estimates labeled "SR" are for candidates who participated in self-reflection, while estimates labelled "C" are for candidates who participated in coaching.