EdWorkingPaper No. 23-845

"Refining" Our Understanding of Early Career Teacher Skill Development: Evidence From Classroom Observations

Brendan Bartanen
University of Virginia

Courtney Bell
University of Wisconsin

Jessalynn James
TNTP

Eric S. Taylor
Harvard University

James H. Wyckoff
University of Virginia

Novice teachers improve substantially in their first years on the job, but we know remarkably little about the nature of this skill development. Using data from Tennessee, we leverage a feature of the classroom observation protocol that asks school administrators to identify an item on which the teacher should focus their improvement efforts. This "area of refinement" overcomes a key measurement challenge endemic to inferring from classroom observation scores the development of specific teaching skills. We show that administrators disproportionately identify two teaching skills when observing novice teachers: classroom management and presenting content. Struggling with classroom management, in particular, is linked to high rates of novice teacher attrition. Among those who remain, we observe subsequent improvement in these skills.

VERSION: September 2023

Suggested citation: Bartanen, Brendan, Courtney Bell, Jessalynn James, Eric S. Taylor, and James H. Wyckoff. (2023). "Refining" Our Understanding of Early Career Teacher Skill Development: Evidence From Classroom Observations. (EdWorkingPaper: 23-845). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/wykj-e663

"Refining" Our Understanding of Early Career Teacher Skill Development: Evidence From Classroom Observations

Brendan Bartanen[1], Courtney Bell[2], Jessalynn James[3], Eric S. Taylor[4], and James H. Wyckoff[1]

[1]University of Virginia

[2]University of Wisconsin

[3]TNTP

[4]Harvard University

Author Note

We thank the Tennessee Department of Education and the Tennessee Education Research Alliance for providing data access. We are also grateful to Julie Cohen and Jason Grissom for their helpful feedback on earlier drafts. Generous financial support was provided by the Spencer Foundation. (#201900117)

Abstract

Novice teachers improve substantially in their first years on the job, but we know remarkably little about the nature of this skill development. Using data from Tennessee, we leverage a feature of the classroom observation protocol that asks school administrators to identify an item on which the teacher should focus their improvement efforts. This "area of refinement" overcomes a key measurement challenge endemic to inferring from classroom observation scores the development of specific teaching skills. We show that administrators disproportionately identify two teaching skills when observing novice teachers: classroom management and presenting content. Struggling with classroom management, in particular, is linked to high rates of novice teacher attrition. Among those who remain, we observe subsequent improvement in these skills.

"Refining" Our Understanding of Early Career Teacher Skill Development: Evidence From Classroom Observations

Introduction

Research on teacher development has struggled to identify rigorous evidence of how to improve teaching skills. This is particularly problematic in the development of novice teachers. While some teachers enter the profession well-prepared to provide high-quality instruction to students, others are much less successful (Atteberry et al., 2015; Kane et al., 2008), leading to worse outcomes for students and high rates of novice teacher attrition (Gray & Taie, 2015; Redding & Nguyen, 2020; Smith & Ingersoll, 2004). For example, the top quartile of teachers entering Tennessee classrooms outperform the bottom quartile, on average, by more than 0.40 SD in student achievement growth (Figure 1, Panel B) and by more than 2 SD when measured by standardized classroom observation scores (Figure 1, Panel A).

However, substantial empirical evidence documents that teacher effectiveness improves with experience. Much of this returns-to-experience work examines student test scores, showing that, on average, novice teachers meaningfully improve their value-added to student achievement over the early years of their careers (Atteberry et al., 2015; Boyd et al., 2008; Clotfelter et al., 2007; Harris & Sass, 2011; Kraft & Papay, 2014; Ost, 2014; Rivkin et al., 2005; Rockoff, 2004). More recent studies replicate this early career improvement as measured by rubric-based classroom observation scores (Bell et al., 2023; Kraft et al., 2020). Importantly, early career improvement—for both value-added and observation scores—is driven by novice teachers who are initially the lowest performing (Figure 1). These two empirical results—substantial variability in novice teacher effectiveness at entry and the negative relationship between initial performance and returns-to-experience growth—offer an opportunity to learn more about how to reduce early career performance differences and understand the nature of teacher skill development, more broadly. In particular, these two results invite questions such as: What skills do early career teachers improve? What are the conditions that lead to improvement? Existing research provides only limited insights to these questions; a better understanding of which could have meaningful implications for teacher preparation, teacher induction, professional development, and the ability of teachers to improve student outcomes.

What we do know is that early career or novice teachers learn to teach in varied organizational contexts, including preservice teacher education programs, local induction programs, and a wide range of professional development programs. Researchers have analyzed how teachers develop teaching skills in these contexts for decades (for a recent review of teacher learning writ large, see Russ et al., 2016). Recent analyses of thousands of classroom observations conducted with multiple observational measures suggest at least two core domains of teaching skills can be measured—general instructional methods and classroom management (Ferguson & Danielson, 2015). However, important questions about how teachers learn those domain skills remain unresolved. For example, are some skills "foundational" in the sense that some minimum level of competency is necessary before other skills flourish (Kagan, 1992); or are skills learned interactively where the learning that supports the development of one skill also assists in the development of another skill (Grossman, 1992)? Unfortunately, evidence on these and a range of other questions regarding teacher development is sparse and often based on anecdotes or small case studies where competing hypotheses are rarely addressed.

In this paper, we employ teacher observation data for Tennessee teachers to better understand which skills contribute to early career teacher development. We leverage a feature of the Tennessee classroom observation system in which the administrator observing a teacher is required to identify one of 19 observed skills that the administrator believes would be most influential for the improvement of the teacher. This "area of refinement" provides an opportunity to understand how teaching skills interrelate and contribute to teacher development. Using data on roughly 25,000 novice teachers, we find that classroom observers disproportionately identify two skills relevant for the improvement of the lowest-performing novice teachers: classroom management and presenting content. We show that this finding likely reflects real differences in these skills (i.e., low-performing novice teachers have relatively weaker skills in these areas) combined with administrators' judgements of skill development necessary to become a successful teacher. We also show that the improvement of teachers' overall observation scores or value-added is linked to the improvement of these administrator-identified skills. Finally, these results are unlikely to be driven by preconceived notions of administrators, teacher attrition, or teacher sorting.

To our knowledge, this is the first analysis to employ large-scale quantitative data to investigate specific skills associated with the overall development of novice teachers. We help to bridge two largely parallel literatures examining, respectively, the returns to teacher experience and the development of novices' teaching skills. Our contribution comes from leveraging the "area of refinement" (AOR) feature of the Tennessee teacher observation protocol. The AOR addresses a common measurement issue in teacher observations in which most items and domains of the rubric are highly correlated (e.g., Kane & Staiger, 2012), limiting their ability to provide insight about teacher skill improvement. Asking administrators to identify a single focus area for a teacher's improvement efforts mechanically removes this correlation and capitalizes on their insider expertise. We demonstrate how administrators disproportionately identify classroom management and presenting content as the AOR for low-performing novice teachers. Struggling with classroom management, in particular, is linked with high rates of attrition among novice teachers. Among those who remain, we find subsequent improvement in classroom management and presenting content skills.

Background

Improving measured teaching performance and teachers' contribution to student outcomes has been a longstanding goal of education policy. As a result, there is a vast literature that addresses different aspects of the goal. Teachers develop teaching skills through a variety of experiences, some formal, some informal, some prior to entry into teaching, and some during the early years of their careers. Educational policymakers and practitioners design these experiences with the goal of cultivating the skills thought to be critical to improving overall teaching quality and student outcomes. We divide this substantial literature into three related questions.

- To what extent do policies or practices affect teacher development and student outcomes?
- How do the attributes of teachers, including their teaching experience, and their schools influence their development?
- How do individual teaching skills develop and does the dynamic acquisition of these skills affect the overall development of teaching quality?

In this paper, we are particularly interested in literature related to the third question on skill development among early career teachers, but acknowledge that our work is informed by teachers' skill development more generally. As we summarize below, our work builds on and is situated in the much broader theoretical and empirical literature on professional development. We briefly summarize this research, especially as it relates to skill development by early career teachers. We then examine in more detail the literature related to the development of specific teaching skills. Finally, we review prior work on the returns to teacher experience, which is almost exclusively focused on generalized measures of effectiveness, such as value-added and average classroom observation score. This review frames our main contribution, which is to connect skill development to the returns to teacher experience using a different feature of Tennessee's classroom observation protocol.

Effective professional development

Research on teacher professional development across the career continuum finds uneven impacts. Somewhat older studies find that professional development can improve measured teaching quality or student outcomes but often does not (see, for example, Garet et al., 2008; Garet et al., 2016; Garet et al., 2011; Kennedy, 2016; Yoon et al., 2007). More recent meta-analyses find more consistent average positive impacts on measured teaching quality (c.f., Garrett et al., 2019) and on student outcomes (Lynch et al., 2019), but the impacts are heterogeneous. Most of these studies review professional development interventions that are offered by experts employed outside of typical preservice, school, and classroom contexts. While teachers regularly participate in externally initiated or structured "interventions" (Wei et al., 2010), there is a great deal or learning that can occur interacting with colleagues that are employed inside of those contexts in "business as usual" interactions. These colleagues may also be experts, but the interactions with teachers are a part of the schools' internal routines. Teachers spend most of their time developing and teaching in these routine contexts. Therefore, it is important to understand teacher development in routine or "business as usual" contexts.

Evidence suggests that professional development efforts embedded in the context of actual routine school and classroom processes can be effective. For example, the most compelling evidence on teacher preparation suggests that high-quality clinical experiences improve novice teacher effectiveness and the outcomes for the students they teach. Further, clinical experiences in which candidates are matched to skilled mentor teachers and in schools with strong professional learning environments can meaningfully improve candidates' teaching skills (for a good summary of this research, see Ronfeldt, 2021).

Similarly, teacher coaching is perhaps the most effective intervention to improve teaching skills and student outcomes. A meta-analysis of rigorous studies finds that teacher coaching can substantially improve instruction and student achievement, but the effects are much smaller when coaching is conducted at scale (Kraft et al., 2018). Assessing teaching skills with tailored feedback was intended to be a principal component of teacher evaluations embedded in several Obama-era federal education reforms, like Race to the Top and the Teacher Incentive Fund. When these efforts included assessments that rigorously differentiated teaching effectiveness and meaningful feedback or coaching, they improved instruction and student outcomes (Dee & Wyckoff, 2015; Taylor & Tyler, 2012). This evidence suggests that effective professional development can also be carried out by internal actors who are embedded in daily ongoing teaching contexts, assessing individual strengths and weaknesses, and providing individual or small group feedback or coaching.

The development of specific teaching skills

The development and measurement of teachers' professional skills and expertise has been an area of inquiry for decades (for recent reviews, see Cappella et al., 2016; Russ et al., 2016). Recent literature depicts the development of teaching skills as multi-dimensional, situated, and dynamic. Teachers are not simply effective or ineffective—there are many instructional skills they need to develop, and this development may occur in different ways (Malmberg et al., 2010). The specific skills identified by researchers as relevant to effective teaching vary; however, rich literature reviews and analyses of thousands of classroom observations conducted with multiple observation systems suggest at least two core domains of teaching skills can be measured: instructional methods and classroom management (Ferguson & Danielson, 2015; Hafen et al., 2015; Kane & Staiger, 2012). Teachers need instructional skills such as effectively conveying content, getting students to think deeply about their own learning, and providing responsive feedback, but they also need classroom management skills so they can successfully engage groups of children and adolescents in meaningful learning. The evidence supporting the importance of these skills to student outcomes is strong and rigorous (see, e.g., Cappella et al., 2016, for a summary). Less clear is how teachers build these skills.

The teacher education literature includes conjectures on the best way to develop teaching skills and the dynamic nature of skill development. Some researchers have argued that development occurs in sequential stages, with deeper learning of more advanced pedagogical skills following the development of classroom management skills (e.g., Fuller &

Bown, 1975; Kagan, 1992). Others have questioned this stage-based approach arguing that skills develop interactively and incrementally (Feiman-Nemser et al., 1999; Grossman, 1992; Watzke, 2007). The evidence employed to support these theories comes from case studies and small-scale qualitative analyses documenting the practices of early career teachers and their self-assessment of the challenges they confronted. None include measures of teacher performance or student outcomes or rigorous designs that could quantify the effects of different approaches to overall skill development. Nonetheless, both perspectives remain influential in teacher education.

A recent review of the literature concludes by posing questions unresolved by research: "What is the skill trajectory of a teacher's ability to engage students, teach academic skills, or manage behavior? What accounts for variation in these skill trajectories?" (Cappella et al., 2016). Another review summarizes skill development as evolving through stages that begin with the acquisition of fact-like knowledge that is subsequently transformed into usable knowledge through pedagogical reasoning and the building of routines. This routinization of knowledge occurs as teachers gain classroom experience (Russ et al., 2016). Whether the development of effective teaching requires sequential skill development (at least to some threshold) or whether skills develop simultaneously and interactively remains an empirical question. Regardless, there is robust evidence that both instructional skills and classroom management skills are necessary to effective teaching.

Yet, there is a strong sense among entering teachers, their principals, and researchers who study novice teacher learning that entering teachers are frequently ill-prepared to manage their classrooms. An inability to effectively manage student behavior is among the most frequently cited reasons by early career teachers leaving their schools (e.g., Dickenson et al., 2022; Kwok, 2021; Robertson, 2006). In an influential review of teacher preparation, Levine (2006) finds that only a third of surveyed principals thought that schools of education prepared teachers to "maintain order and discipline in the classroom" either very well or moderately well (Levine, 2006, p. 33). In contrast, principals believed entering teachers possessed much stronger skills on mastery of their subject area (72% for moderately or very well), understanding how students learn (54%), and utilizing different pedagogical approaches (54%).

Studies of teacher preparation suggest that, historically, many preparation programs were not attentive to designing strong student teaching experiences or building effective classroom management skills (e.g., Freeman et al., 2014; Greenberg et al., 2014; Hammerness, 2011; Stevenson et al., 2020), although emerging evidence suggests recent improvement (Pomerance & Walsh, 2020). Prior work has focused on documenting both the requirement or actual frequency of classroom management coursework and the perceived quality of this training. For instance, Hammerness (2011) found that fewer than half of traditional teacher preparation programs in New York City required any coursework in classroom management. Based on 2015 data from the National Council for Accreditation of Teacher Education, fewer than 15% of preparation programs required classroom management for all licensure candidates (Stevenson et al., 2020). The lack of explicit coursework may betray the overall prevalence of classroom management training, however. In a 2014 analysis, the National Council on Teacher Quality (NCTQ) described classroom management training in teacher preparation as "everywhere but nowhere" (Greenberg et al., 2014). Specifically, they found that, among surveyed programs, nearly all mentioned classroom management in curriculum documents, but the actual training was often scattered throughout the curriculum, not consistently evidence-based, and did not require the actual *practice* of classroom management skills.

More recent work suggests some improvement. An updated NCTQ analysis notes an uptick in the percentage of programs teaching evidence-based classroom management strategies (Pomerance & Walsh, 2020), though many programs are still graded poorly in this area. Other recent commentaries echo that while the prevalence and intensity of classroom management training has increased in recent years, there remains much work to be done in terms of implementing evidence-based practices and successfully integrating classroom management training throughout the preparation curriculum and clinical practice experiences (Kwok, 2021; Stevenson et al., 2020).

The returns to teacher experience

Substantial empirical evidence documents that teachers' abilities to improve student achievement improve with experience, on average. These return-to-experience analyses, which have been replicated across several states, different achievement tests, and time periods, find that the typical novice teacher meaningfully improves their ability to raise student achievement over the first four to seven years. Specifically, teachers are estimated to improve their value-added to student achievement by 0.05–0.12 student-level standard deviations over the first five years of teaching (Atteberry et al., 2015; Boyd et al., 2008; Clotfelter et al., 2007; Harris & Sass, 2011; Kraft & Papay, 2014; Ost, 2014; Rivkin et al., 2005; Rockoff, 2004).

However, this average relationship masks substantial variability in novice teachers' initial effectiveness and subsequent improvement. Atteberry et al. (2015) estimate that the bottom quintile of first-year New York City teachers were more than 0.20 SD (in terms of student test scores) less effective than the average first-year teacher. This bottom quintile improved by more than 0.15 SD, on average, over the first five years of teaching. By contrast, the highest quintile in terms of initial performance showed minimal evidence of improvement with experience. Using data Charlotte-Mecklenburg Schools, Kraft and Papay (2014) find this same pattern of heterogeneity, where the bulk of improvement is driven by novice teachers who are initially the lowest performers.

Can classroom observations help us understand how skill development is related to improvements in the early years of teaching?

Importantly, the returns to experience literature documents that the typical novice teacher is improving over time on summative outcomes, but it does not tell us *what* is improving or *how* the novice teacher improves. Without that knowledge, it is difficult to develop policies that could benefit novice teachers and their students. The increasing use and availability of rubric-based classroom observation scores present a potentially promising empirical window into skill development among novice teachers. Instead of a summative, outcome-based measure like value-added, observations aim to identify performance on various dimensions purported to capture effective teaching, such as lesson planning, checks for understanding, and classroom management.

Recent studies replicate the returns to experience findings (Bell et al., 2023; Kraft et al., 2020; Papay & Laski, 2021), again demonstrating both that the average novice teacher improves substantially over their first five years and that this improvement is driven by teachers who are initially the lowest performing. In considering whether classroom observation can accurately reflect improvement in teacher performance, reasonable concerns exist regarding rater leniency and bias, the effect of incentives tied to the observations, or the manipulation of ratings by teachers unrelated to their true performance (e.g., Bell et al., 2023; Kraft & Gilmour, 2017). Any of these issues might create the perception that teacher performance had improved well beyond improvements in actual performance and, as such, would threaten the validity of findings that use improvements in observation scores to better understand differential improvement in specific teaching skills. Employing classroom observation data from Tennessee and the District of Columbia, Bell et al. (2023) find that the improvement of teachers' overall classroom observation scores over the early years of their careers can be interpreted as the causal effect of experience on job performance, as measured by the specific teaching practices included in the rubric. While this finding does not guarantee that the scoring of specific practices overcome some of the aforementioned concerns, it does increase our confidence in the validity of the observation process, particularly in Tennessee and DC.

As is common in many classroom observation systems implemented at scale with administrators rating teachers, Kraft et al. (2020) find the eight domains included in the

Charlotte-Mecklenburg observation rubric all load strongly onto one principal component. The Measures of Effective Teaching (MET) study finds essentially the same result (Kane & Staiger, 2012), particularly for the Framework for Teaching instrument, which is similar to the rubrics employed in Charlotte-Mecklenburg and the current study. It is unclear from these analyses whether the design of the rubric does not capture conceptually distinct domains or whether raters fail to implement the rubric faithfully. Regardless, this effectively precludes employing the domain-level data to explore the development of different skills, as domain or item-specific scores move in tandem. Even so, it is worth noting that a larger share of novices scored “Below Standard” or “Unsatisfactory” for management of student behavior (3.5%) and instructional presentation (2.3%) than any of the other domains (Kraft et al., 2020).

The current study builds upon the existing literature by leveraging a feature of Tennessee’s classroom observation protocol in which administrators (raters) identify a single item on which the teacher should focus their improvement efforts. As we demonstrate in the remainder of the paper, this “area of refinement” allows us to interrogate how the development of specific teaching skills may be related to some teachers’ improvements over the first five years of their careers.

Data and Measures

Data

We access statewide longitudinal data from Tennessee via the Tennessee Education Research Alliance and the Tennessee Department of Education. These data include detailed information on students, educators, and schools and are first available in 2001–02. For the current analysis, we focus on the 2012–13 through 2018–19 school years; as discussed below, this period covers the years following Tennessee’s teacher evaluation reforms. Appendix Table A1 shows descriptive statistics for the analytic sample of teachers. Important for this analysis, the data files allow us to observe when new teachers enter the K–12 public education system in Tennessee and their subsequent job history, including detailed information from the evaluation system. The evaluation data include information from roughly 650,000 classroom observations conducted over the period, including item-level scores (i.e., a 1 to 5 rating for each assessed rubric item), an identifier for the rater (nearly always a principal or assistant principal), and the item chosen by the rater as the "area of refinement." We describe these key data elements below.

Measures

Beginning in the 2011–12 school year, Tennessee implemented a high-stakes educator evaluation system that requires teachers to undergo multiple rubric-based classroom observations conducted by their principal or assistant principal. The number of observations typically ranges from 1 to 4 and is largely dependent on teachers' years of experience and prior performance (Hunter, 2018). As shown in Appendix Figure A1, teachers with fewer than three years of experience typically receive four observations, though we also observe small percentages who receive fewer than that. Among teachers with three of more years of experience, roughly 60% receive two observations and 30% receive one observation. Roughly 80% of teachers in Tennessee are evaluated using the TEAM rubric. The rubric covers three domains (*Instruction*, *Planning*, and *Environment*), which collectively contain 19 items (the complete rubric is shown in Appendix C). For a given observation, teachers are typically assessed on two or three domains: *Instruction* and one or both of the other domains. In addition to providing discrete 1 to 5 ratings on each assessed rubric item, raters also select one item as the "area of refinement," (AOR) which is defined by the TDOE as an "area in which the [rater] needs to help the teacher improve." According to TDOE guidelines, the AOR should be the focus of the post-conference discussion between the teacher and rater.

How do raters choose the area of refinement?

In the TDOE evaluation handbook, raters are encouraged to consider the following criteria in selecting the AOR: (1) Which areas on the rubric received the lowest scores? (2) Which of these areas would have the greatest impact on student achievement or other areas of the observation rubric? (3) In which area will the teacher have the most potential for growth? The handbook also encourages raters to choose an AOR for which there is sufficient and specific evidence from the lesson to support the choice. While there is no guarantee that raters follow this guidance, it likely does provide some insight into raters' decision-making process. Based on these criteria, we propose that that the AOR is a function of two factors. The first is a teacher's relative skills. That is, the AOR will tend to be an item on which the teacher is lower performing (relative to other items). The second factor is salience, which concerns the rater's judgment about the relative importance of items for particular goals, such as improving student outcomes. The salience factor implies that even if a teacher is equally skilled across dimensions captured by the observation rubric, certain skills may be so fundamental for (improvements in) effective teaching or that they are chosen as the AOR even if they are not a teacher's weakest area.

Measuring teacher skill development using the area of refinement

Because the AOR choice partially reflects item salience, we should be careful in using it to make inferences about teachers' skill development. Specifically, changes over time in item frequency as the AOR could indicate changes in a teacher's relative skills, but they could also reflect changes in salience that are correlated with teacher experience. For example, raters may believe a priori that novice teachers should focus on improving classroom management skills. Even if classroom management is not among a novice teacher's weakest skills, we still may observe higher frequency of this item as an AOR as compared to veteran teachers. Similarly, if classroom management becomes less frequent as the AOR as teachers gain experience, this could conflate skill development (i.e., novice teachers improving their classroom management skills) with decreased salience (i.e., raters' judgments about the importance of focusing on improving classroom management decreases). In our empirical work, we help to elucidate what drives changes in the AOR.

Area of refinement versus item scores

For measuring teachers' relative skills, there are several reasons why the AOR might provide better information than the item scores. The first reason concerns the stakes. Different from the item scores—which are inherently high-stakes as they are used to construct the teacher's average observation score that receives 50% or more weight in the summative evaluation rating—the AOR has no formal stakes other than its role as one of many items. This absence of stakes may resolve a fundamental tension in teacher evaluation systems between the dual purposes of evaluation and development, whereby raters are hesitant to assign low overall ratings to teachers (particularly novices) because of the potential consequences (Grissom & Loeb, 2017). Because the AOR has no bearing on the teacher's summative evaluation, there is no obvious reason for raters to deviate from the goal of providing accurate and actionable feedback to teachers.

A second reason why the AOR might contain better information than other item scores is the cognitive nature of the task. Choosing the AOR is much simpler than assigning ratings to each of 19 individual items—raters only need to determine a rank ordering. Particularly when there are many items, asking raters to accurately assess each of them in the context of a single classroom observation is cognitively demanding. Additionally, raters might not even be able to observe sufficient evidence to provide a judgment on all items, depending on the lesson. Each of these may lead to measurement error in the non-AOR individual scores. Prior empirical evidence in several contexts demonstrates that item scores tend to be very highly correlated (e.g., Bartanen & Kwok, 2021; Kane & Staiger, 2012; Kraft et al., 2020), suggesting that raters do not reliably distinguish among teachers' skills in different areas when assigning scores. While identifying the specific drivers of this pattern is beyond the scope of this paper (and what we can accomplish with large-scale administrative data), we suggest that the unreliability in the item scores may make them poorly suited for understanding how novice teachers develop skills in particular areas.

A final benefit of the AOR is that it implicitly incorporates information about the relative importance of different teaching skills. Whereas the item scores receive equal weight on the observation rubric and raters must assign a 1 to 5 rating to each item (if the parent domain is assessed on that observation), the AOR process leverages the rater's insights for how this teacher could improve pedagogy and student outcomes. As an example, *Managing Student Behavior* (one of the items on the TEAM rubric) may be a particularly important or fundamental skill for effective classroom instruction. However, its weight in determining a teacher's overall observation score is quite low because it is just one of 19 items and, further, the *Environment* domain is only assessed once or twice per year for the typical novice teacher. By asking raters to identify just a single item, the AOR process effectively prioritizes skills that (raters believe) are most critical.

Methods

Our primary goal is to use the area of refinement to understand novice teacher skill development. Generally speaking, our approach compares AOR frequency across observation items among novice teachers to provide insight on teacher development. Below, we operationalize the AOR as a measure of teachers' skills. We then describe our estimation approach.

Operationalizing area of refinement indicators

As described earlier, for each classroom observation raters are required to choose one rubric item as the AOR. In addition to conceptual considerations, there are several empirical challenges to leveraging these indicators as measures of teachers' (relative) skills. First, because an AOR is recorded for each observation, teachers who received more observations in a given year also have more identified AORs. Thus, a simple comparison of novices and veterans on a particular item (i.e., whether a teacher flagged for refinement on the item at least once during the year) is misleading because novices tend to have more observations. As described below, our approach estimates the novice–veteran difference at the level of the *individual observation*, rather than at the teacher-by-year level.

Second, because raters must select exactly one AOR from the items considered on a particular observation, the baseline likelihood of a teacher being flagged on a particular item is a function of how many other items could have been chosen. For example, if a teacher is observed only on the *Planning* domain (four items), the simple probability of being flagged on any item is 0.25. If the teacher was observed on *Instruction* (12 items) and planning, this simple probability drops to 0.06. Additionally, if raters are selecting the AOR based on a relative ranking of the teacher's skill among observed items, the relative ranking for a given item depends on which other items are being observed. For example, if a teacher's worst skill is classroom management, their likelihood of being flagged for refinement on *Planning* or *Instruction* domain items is lower when the observation includes the *Environment* domain (above and beyond the change in probability due to more items being observed).

To overcome these challenges, our primary analyses treat teacher-by-year-by-observation as the unit of observation and we implement a weighting scheme that adjusts for the number of observations over the year and the number of items considered during a particular observation. For a given observation j for teacher i in year t, the weight is:

$$\frac{n_{jit}}{\sum_{j=1}^{J} n_{jit}} \tag{1}$$

where n is the number of items that were rated during observation j and J is the total number of observations teacher i received in year t. Effectively, the total weight for a teacher-by-year is 1.0, and an individual observation receives greater weight when more items were rated. The intuition of this weighting is that when a particular item is chosen as the AOR over a larger number of alternative items, we should give it greater weight (and vice-versa for when there are fewer items). When *Managing Student Behavior*, for instance, is chosen in an observation where all three domains were assessed (*Instruction*, *Planning*, and *Environment*), that choice provides a stronger signal about the teacher's relative skills than in an observation where *Environment* was the only assessed domain. In the latter case, the choice of *Managing Student Behavior* contains less signal because the rater only chooses among four items, instead of 19.

When using an AOR indicator as an outcome—for instance, predicting the difference between novice and veteran teachers in being flagged for the *Questioning* item—we also need to account for the fact that unless the teacher was actually rated on the relevant domain (*Instruction*, in this case), the outcome is undefined. Thus, when predicting the AOR for *Questioning*, we should not include observations where the *Instruction* domain was not assessed. Here, we can simply construct the weight in Equation 1 for each of the three domains, where the denominator only sums over observations where the given domain was assessed. In the aforementioned example of using *Questioning* as the outcome, we use the weight constructed using observations where instruction was assessed. We call these "domain-specific weights." Appendix B includes more details about weighting and a comparison between weighted and unweighted estimates.

Modeling

Our primary aim is to understand whether novice teachers are flagged for refinement in some areas more than others and, the extent to which there are differences between lower-performing and higher-performing novices. To do so, we employ the weighted frequencies described above. For most raters the choice of AOR reflects both relative skills and salience. For example, *Questioning* is the most frequently selected item for an area of refinement. Its frequency likely reflects multiple factors, including: (1) the

Questioning "skill" is relatively difficult, (2) the standards defined by the rubric for *Questioning* make it difficult to earn high scores, and (3) raters are more likely to choose this item because they believe it is important (e.g., for improving student outcomes) or because its manifestation in a lesson is more obvious.

We can tease apart some of the dynamics that drive variation in item frequency by comparing novice and veteran teachers. For instance, if certain items are more frequent because of how they are defined in the rubric or because they are more salient to raters, we should observe this for both novice and veteran teachers. Comparing items across novice and veteran teachers, then, provides an indication of which skills novices, in particular, struggle with. Thus, we are effectively comparing the predicted probability that a particular item is chosen as the AOR across three groups: low-performing (Q1) novices (a first-year teacher who is in the bottom quartile of average observation scores among all first-year teachers), higher-performing (Q2–Q4) novices, and veterans (teachers with 10+ years of experience). To this end, we estimate via least squares regression:

$$\text{AOR}_{ijt} = \delta(f(\text{Experience}_{it})) + \tau_t + \text{Domains}_{ijt} + \epsilon_{ijt} \tag{2}$$

where AOR_{ijt} is a binary indicator for whether a given rubric item (e.g., *Questioning*) was chosen as the area of refinement for teacher i in observation j in year t. We parameterize teacher experience ($f(\text{Experience}_{it})$ as a set of indicator variables for each specific year of experience up to 9 years, then a 10+ year bucket. However, we further separate the first three years of experience into separate buckets for Q1 and Q2–Q4 for novices, as defined above. We include fixed effects for year (τ_t) to adjust for state-level secular trends in the frequency of refinement areas. We also include fixed effects for each combination of rubric domains (Domains_{ijt}) that were assessed in a given observation (e.g., *Instruction* only, *Instruction* and *Planning*, etc.), which accounts for variation in the total number of items that a rater could choose for the AOR. δ are the coefficients on the experience buckets, which in this model represent the difference in probability that the given item will be flagged as the AOR for teachers at differing levels of experience (by initial performance).

Estimated differences in refinement areas from Equation 2 could reflect multiple mechanisms. If, for example, novice teachers are more likely to be flagged for refinement in *Managing Student Behavior*, this could be driven by improvement (i.e., a teacher becomes less likely to receive this item as the AOR as they gain experience), but also because of differential attrition and/or non-random sorting to particular school environments. Under differential attrition, novice teachers who are flagged for *Managing Student Behavior* leave the profession at higher rates, which lowers the observed frequency of this item among veteran teachers, even absent actual improvements in this skill. Under non-random sorting, teachers transfer to schools where *Managing Student Behavior* is easier (or at least less likely to be chosen as the AOR), again creating an experience gradient.

To isolate improvement from these alternative mechanisms, we can add to the model teacher and school fixed effects. The inclusion of teacher fixed effects effectively makes Equation 2 a within-person, returns to experience design: the marginal effect of experience is identified by comparing teachers to themselves across time.[1] While this approach avoids bias from selective attrition in the estimated returns to experience, it also introduces challenges due to the collinearity of experience and year. Specifically, because most teachers gain a year of experience with each additional calendar year, we cannot separately estimate coefficients for experience and year fixed effects without additional

[1] Including teacher fixed effects also introduces a potential external validity concern, whereby the estimates at a particular experience level are identified only from teachers who remained in the profession for that length of time. One might reasonably believe, for instance, that teachers who left the profession early in their career would not have experienced the same amount of growth as teachers who persisted. We can shed light on this possibility by examining the early career returns to experience for subsets of teachers based on how long they remain. For instance, we can compare the year 0 to year 1 growth among teachers who remained for 2, 3, etc., years. If growth is similar across these sub-samples, it provides suggestive evidence that the teacher FE results are likely to be externally valid for the subset of novice teachers who left the profession early. We carry out this exercise in Appendix Figure A2 for all novice teachers and Q1 novice teachers, respectively. Though power is limited, particularly for the Q1 analysis, we find little evidence to suggest that the returns to experience (based on the AOR) differ across these sub-samples. Therefore, we believe it is likely that the teacher FE results provide a reasonable estimate for the returns to experience for the full sample of novice teachers, not just those that persisted.

assumptions. As described above, our choice to group teachers with 10 or more years of experience allows us to avoid this collinearity problem. This choice introduces an implicit assumption that the marginal effect of additional experience is zero once a teacher reaches their 10th year, which appears to hold in our data.[2] Using this veteran teacher group, we can estimate coefficients for the year fixed effects, which are then applied to teachers with 0–9 years of experience.

Results

We begin by establishing two key empirical results. First, novices and veterans differ substantially in which rubric items are chosen by raters as the AOR. Specifically, novices are substantially more likely to receive an AOR relating to classroom management and presenting content skills, whereas veterans are more likely to receive an AOR relating to higher-order teaching skills, such as implementing activities that teach students problem-solving skills. Second, these novice–veteran AOR differences are driven by the lowest-performing novice teachers; higher-performing novices and veterans are similar across most items. Our interpretation of these results is that the sharp improvement in overall effectiveness of novice teachers in their first few years on the job is driven by improvements in fundamental teaching skills. By the same token, the lack of consistent improvement among initially higher-performing novice teachers as well as veteran teachers—who already have minimum competency in these skills—is explained by lack of improvement on higher-order teaching skills.

Area of refinement differences between novice and veteran teachers

Novices and veterans receive different AORs, on average. These patterns are demonstrated in Panel A of Figure 2, which plots the frequency (i.e., the predicted

[2] Specifically, we can examine the marginal effect in the years immediately below the cutoff. To the extent that the experience profile is already quite flat in this region, we are more confident in the assumption of no systematic improvement beyond 10 years.

likelihood of each AOR obtained after estimating Equation 2 for each of the 19 possible AORs between novice (no prior experience) and veteran (10+ years of prior experience) teachers. We further differentiate novices according to their initial performance as measured by average observation score. The predicted probabilities represent, specifically, the likelihood that the item is chosen as the teacher's AOR in an observation that includes all 19 items. From this plot, we can understand which items are more commonly identified as an AOR, overall, and which items have a large novice–veteran gap. We also plot the corresponding differences for the item scores in panel B. For parsimony, Figure 2 omits confidence intervals and tests of statistical significance (see Appendix Table A2 for this information), though we note that the analyses are highly powered given the large sample size. To facilitate comparison, we order the items according to the size of the predicted difference between Q1 novices and veterans. We also include the alpha-numeric identifier for each item (e.g., "E2" for *Managing Student Behavior*), where the letter denotes the rubric domain (*Instruction*, *Planning*, or *Environment*) and the number denotes the order within the domain (e.g., "E2" is the second item listed in the *Environment* domain).

Two items emerge with substantial differences between novices and veterans: *Managing Student Behavior* and *Presenting Instructional Content*. For the former, Q2–Q4 novices are roughly three times more likely than veterans to be flagged for refinement in this area (5.6% vs 1.8%), while Q1 novices (11.7%) are six times more likely than veterans. Each pairwise difference is statistically significant ($p < 0.001$). Other items on the *Environment* domain of the rubric are rarely chosen as the AOR for veterans or novices, demonstrating the salience of *Managing Student Behavior*. We observe similar disparities for *Presenting Instructional Content*, with an AOR probability of 4.9% for veterans, 7.2% for Q2–Q4 novices, and 14.7% for Q1 novices ($p < 0.001$ for each pairwise difference).

The structure of the AOR process effectively requires that if novices are more likely to be flagged for refinement in some areas, they are less likely to be flagged in others. This tradeoff is not equally distributed among all items, however. We observe a decrease in

AOR frequency among items that represent higher-order teaching skills. Figure 2 demonstrates this; we find that Q1 novices are less likely to be flagged for refinement in several areas, including *Questioning*, *Feedback*, *Grouping Students*, *Thinking*, *Problem-Solving*, and *Assessment*. Again, Q2–Q4 novices tend to be more similar to veterans than Q1 novices. For several items, there are no differences in AOR frequency for higher-performing novices and veterans, including *Questioning* (the most frequent AOR item) and *Standards and Objectives*.

Comparing panels A and B of Figure 2 demonstrates how the AOR provides different information than the item scores. Whereas the novice–veteran gaps for the AOR vary substantially across items, the item score gaps do not. There are some differences across items in terms of their mean ratings—patterns we return to later—but the gaps between veterans and Q1 and Q2–Q4 novices, respectively, are very similar. Simply comparing novices to veterans across the item scores would suggest that no particular skills drive the large effectiveness differences between these groups. As has been demonstrated in a variety of contexts using these types of classroom observation rubrics, item scores tend to be very highly correlated, suggesting that raters do not assign scores that sufficiently distinguish among various dimensions of teaching practice. By contrast, the AOR process facilitates this differentiation and allows raters to identify skills that are perhaps more important for improving the overall quality of instruction.

Overall, the AOR patterns in Figure 2 suggests that novice and veteran teachers differ substantially in their relative skills. Novice teachers tend to struggle with classroom management and basic presentation of instructional content. Veteran teachers are rarely assigned an AOR in these areas; instead, they are asked to focus on improving higher-order teaching skills such as asking and soliciting answers to questions, facilitating high-quality partner and group work, and incorporating opportunities to develop students' higher-order thinking and problem-solving skills. Importantly, these relative skill differences appear to map onto teachers' overall effectiveness levels. That is, a focus on improvement of foundational teaching skills is driven by the lowest-performing novice teachers, whereas higher-performing novices are more similar to veteran teachers.

To make the link between the AOR and teacher effectiveness more explicit, Figure 3 shows results from separate regression models predicting, respectively, value-added and average observation scores in year t as a function of the AOR a teacher receives. Examining value-added, in particular, is helpful because it demonstrates the extent to which the information captured in the AOR maps onto differences in outcomes for students. In each model, which we estimate separately for novices and veterans, we choose *Questioning* (the most frequently chosen item) as the reference category, such that the coefficients for each item indicate the predicted difference in effectiveness compared to a teacher whose AOR was *Questioning*. We also include school-by-rater-by-year fixed effects to account for potential differences in rating standards or school context that are correlated with the likelihood of a particular AOR and the measures of teacher effectiveness. While the models include the full set of 19 items, for parsimony we present results for four focus items: *Presenting Instructional Content*, *Managing Student Behavior*, *Questioning*, and *Problem-Solving*. The full results are shown in Appendix Figure A3.

Overall, we find meaningful relationships between the AOR and both value-added and average observation scores. These relationships are very similar for novices and veterans. Notably, the items with the largest novice–veteran gaps (i.e., items that are more common among novices) in Figure 2 are also those which correspond to the lowest performance. Specifically, we find that novice teachers receiving *Managing Student Behavior* as their AOR have substantially lower observation scores (-0.44 SD) and value-added (-0.14 SD), as do novice teachers receiving *Presenting Instructional Content* (-0.35 SD for observation scores and -0.11 SD for value-added). Similarly, items that were relatively less common as the AOR for novices are associated with higher performance. Specifically, teachers receiving *Problem-Solving* as the AOR area tend to have the highest observation scores. These teachers also tend to have higher value-added, though the differences among most of the items are smaller in magnitude.

Figures 2 and 3 establish, respectively, that raters are more likely to ask novice teachers to focus improvement efforts—via the choice of the AOR—in certain areas like presentation of content and classroom management, and that differences in which AOR a teacher receives are correlated with their overall performance as measured by both average observation score and value-added. As outlined above, these patterns suggest that novice teachers—particularly those with initially lower performance—tend to struggle with specific teaching skills and that their rapid improvement in the first years on the job is driven by skill development in these areas. The remaining analyses aim to support this interpretation. First, we show that the novice–veteran AOR differences largely result from relative skill improvement in *Managing Student Behavior* and *Presenting Instructional Content* and are not merely the result of differential attrition from the profession or nonrandom sorting of teachers to schools. We then examine the assumption that the AOR actually contains information about teachers' relative skills as opposed to other explanations, such as raters' pre-existing beliefs about where novice teachers should improve.

Improvement, attrition, or sorting?

Most of the novice–veteran gap in *Presenting Instructional Content* and *Managing Student Behavior* can be attributed to within-teacher improvements—defined here as a decreased likelihood that these items are chosen as the AOR (Table 1). In addition to these two items, we again include results for *Questioning* and *Problem-Solving*, with results for the full set of items available in Appendix Table A2. Columns 1–3 estimate the gap between Q1 novices and veterans, with columns 4–6 comparing Q2–Q4 novices and veterans. Models (1) and (4) show the baseline gaps presented in Figure 2. The next two models add teacher and school fixed effects. Comparing models (1) and (2), we observe that the baseline gaps of roughly 10 percentage points for *Presenting Instructional Content* and *Managing Student Behavior* shrink somewhat when teacher fixed effects are included.

For *Managing Student Behavior*, the gap shrinks by 37%, with a 21% reduction for *Presenting Instructional Content*. The larger reduction for *Managing Student Behavior* suggests that poor classroom management skills are more strongly linked to attrition among novice teachers, which we show directly in Appendix Table A3. For Q2–Q4 novices, these changes across specifications are smaller in magnitude, which reflects lower attrition rates among higher-performing novices. For *Questioning* and *Problem-Solving*—items which were less likely to be chosen as the AOR among novices—the patterns follow an inverse pattern, as expected.

The teacher fixed effects results demonstrate that novice teachers are relatively less likely to receive an AOR in *Presenting Instructional Content* and *Managing Student Behavior* as they gain experience. The final columns (models 3 and 6) add school fixed effects, which account for potential non-random sorting of teachers to schools. For instance, teachers may seek to transfer to schools where classroom management issues are less prevalent, which could create the observed novice–veteran gap even in models with teacher fixed effects. In general, however, we find limited support for this explanation, as the change in the estimated gaps is fairly small in magnitude.

Figure 4 helps to visualize this pattern of improvement. Specifically, we plot the predicted probability that each of these four items is selected as the AOR as a function of teachers' years of experience and their initial performance (bottom, middle, or top quartiles of first-year average observation score). To maintain the focus on within-teacher improvement, we construct a balanced panel of teachers whom we observe in their first five years on the job. Across each of these items, we observe the likelihood of AOR identification changes much more for the lowest performing quartile and than the remaining 75% of teachers; AOR identification for the highest performing quartile changes very little. Consistent with the results in Table 1, lower-performing novice teachers become relatively more effective in *Presenting Instructional Content* and *Managing Student Behavior* as they gain experience, which is captured by decreasing probability that these items are chosen as the AOR. Correspondingly, there are increases over time in the AOR probability for *Questioning* and *Problem-Solving*, indicating the increasing salience of these items as the focus on *Managing Student Behavior* and *Presenting Instructional Content* declines.

As described earlier, the highest performing quartile of novice teachers show very little improvement the first five years of their careers whereas the lowest quartile novices improve greatly (Figure 1a). Figure 4 provides a potential explanation for this pattern: low-performing novices enter the profession with weak classroom management and presenting content skills, and their improvement is driven by the development of these skills. For higher-performing novices, overall improvement must be driven by higher-level teaching skills, which may be more difficult to develop or for which teachers receive less coaching and support.

Does the area of refinement actually measure teachers' skills?

Our results indicate that raters are much more likely to identify *Managing Student Behavior* and *Presenting Instructional Content* as the areas of refinement for novice teachers than more difficult or higher-level skills such as *Thinking* or *Problem-Solving*, and that these differences are largely determined by the lower performing novices. These patterns appear to reflect both selective attrition (i.e., lower-performing novice teachers leaving the profession at higher rates) and within-teacher improvement. In particular, the improvement patterns are at least consistent with the hypothesis that a key differentiator of novice teachers' initial performance is their skill in areas like classroom management and basic instructional practices.

A key assumption linking the empirical results to our interpretation with respect to early career teacher skill development is that the area of refinement actually captures information about skills. That is, we are assuming that when two teachers receive different AORs, this reflects differences in their relative skills, on average. Similarly, we are assuming that changes over time in which items a teacher receives as the AOR indicates changes in their relative skills. Is this assumption valid?

The novice–veteran AOR gaps presented in Figure 2 could arise for three (not mutually exclusive) reasons. First, they could reflect relative skill differences between novices and veterans. Second, novice–veteran AOR identification differences could reflect a relationship between item salience and the absolute level of performance. Here, novice teachers are more often flagged for *Managing Student Behavior* not necessarily because it is among their weakest skills but because raters are likely to ask teachers to focus on this skill when it is sufficiently low-performing. Baseline proficiency in certain skills, such as classroom management, may be so fundamental to effective teaching that raters will prioritize them over other skills when teachers are lower performing. Finally, novice–veteran differences in the AOR could arise if item salience is directly linked to experience. Under this explanation, a novice and a veteran teacher may have equal performance and skill, but raters believe that identifying *Managing Student Behavior* as an area of refinement is more important for novices than veterans (e.g., the belief that if a teacher has not learned how to effectively manage a classroom over the first ten years of their career, they may be unlikely to do so).

Differentiating among these mechanisms provides insight into the informational content of the AOR. We show that the AOR reflects both relative skill differences and item salience as a function of performance. That is, raters are more likely to ask novices to improve in *Managing Student Behavior* and *Presenting Instructional Content* both because these are more likely to be among novices' weakest skills (in relative terms) and because raters appear to prioritize these items at lower levels of performance (in absolute terms). Importantly, we find little evidence that item salience as a function of experience is a driver of these patterns, which is consistent with the patterns in Figure 3 demonstrating that the correlation between AORs and both average observation scores and value-added were very similar for novice and veteran teachers.

Comparing area of refinement and item scores

As a first step towards understanding the informational content of the AOR, Figure 5 plots a simple bivariate relationship between each item's AOR frequency and the mean score of that item on the 1-5 scale. Note that while item scores are highly correlated (average inter-item correlation = 0.56), which makes it challenging to use them to distinguish among various dimensions of skill development, there is some variation across items in terms of their mean score, which we can use to gain insight about how raters may choose the AOR. We show the relationship between AOR frequency and item scores separately for low-performing novices (Q1), higher-performing novices (Q2–Q4), and veterans—the same groups shown in Figure 2. For each group, we observe, as expected, a negative relationship between the mean item score and the AOR frequency. This relationship is consistent with the claim that the AOR is capturing information about relative skills. In particular, the highest-scoring items, such as *Environment* (E3) and *Respectful Culture* (E4), are almost never chosen as the AOR. Lower-scoring items are, overall, more likely to be the AOR, but there remains substantial variability across items. For all groups of teachers, *Thinking* (I11) and *Problem-Solving* (I12) are among the items with the lowest scores, but they are not particularly common as the AOR. By contrast, *Questioning* (I6) and *Lesson Structure and Pacing* (I4) also tend to have relatively low item scores but are common AORs. These deviations suggest that the AOR also captures information about item salience, which we explore further below.

Focusing on the two items from Figure 2 with the largest AOR gaps between Q1 novices and veterans, *Managing Student Behavior* (E2) and *Presenting Instructional Content* (I3), we find somewhat different patterns. As previously demonstrated in Figure 2, *Managing Student Behavior* is rarely chosen as the AOR for veteran teachers but is very common for low-performing novices. Figure 5 helps explain this gap. For veterans, *Managing Student Behavior* is among the highest-scoring items, at roughly 4.4, on average. This item score is also well above the mean score of 4.1 for veterans. On the basis of item scores, then, *Managing Student Behavior* tends to be among veterans' strongest skills. Although they score absolutely lower on *Managing Student Behavior* (3.9), higher-performing novices (Q2–Q4) perform relatively well on this item and, consequently, it is infrequently chosen as the AOR. This pattern supports the interpretation that novices have relatively weaker classroom management skills. However, among low-performing novices, even though the mean item score for *Managing Student Behavior* (2.8) exceeds the mean overall score, this item is the third most frequent area of refinement highlighting its salience to raters. Further, the substantial difference in AOR frequency between low-performing and higher-performing novices helps to refute the notion that these patterns are merely a manifestation of raters' pre-conceived beliefs about which skills novices should focus on improving. Instead, the AOR gap for *Managing Student Behavior* appears to reflect two things: (1) novices' relative skills in classroom management tend to be lower than veteran teachers', and (2) the salience of classroom management is very high at low levels of performance.

For *Presenting Instructional Content* (I3), the novice–veteran AOR gap also appears to reflect both differences in relative skills and increasing salience at lower absolute performance. Different from *Managing Student Behavior*, however, the relative skill difference on the basis of item scores is much smaller. Veterans and higher-performing novices have item scores that are almost the same as their mean performance, while low-performing novices score slightly lower on *Presenting Instructional Content* than their mean score, on average. Nonetheless, the AOR frequency is much higher for low-performing novices, again suggesting that this item has high salience at lower levels of absolute performance.

Figure 5 provides simple descriptive evidence to suggest that the observed AOR disparities between novices and veterans indicate both differences in relative skills and high salience of particular items at low performance. To further probe this claim, we can make use of the observation-level data. Figure 6 plots results from models that predict the area of refinement in observation j as a function of the discrete 1 to 5 score given in observation j for the subset of four items we examined previously. For each of the five possible scores, we plot the predicted probability that the item is chosen as the area of refinement using both a non-parametric and quadratic fit. We also estimate these relationships separately for novice and veteran teachers, which helps speak to whether raters assign the AOR differently to these groups, conditional on measured performance.

Unsurprisingly, when an item receives a 5, it is almost never chosen as the AOR. As observed in Figure 5, the probability that an item is chosen as the AOR increases at lower item scores, on average, but there are important differences in this pattern across items. For Managing Student Behavior and Presenting Instructional Content, we observe clear non-linearities where the AOR probability at low performance (i.e., a rating of 1 or 2) is very high. For example, when a teacher receives a 1 or 2, respectively, on *Managing Student Behavior*, the probability that this item is chosen as the AOR is 52% and 38%. This probability drops substantially when the teacher receives an "At Expectations" rating of 3 or higher, which suggests that these items lose salience once reaching a higher proficiency level.

For *Problem-Solving*, we find a quite different pattern. The AOR probability remains fairly low even when the teacher receives a 1 or 2, and teachers receiving a 1 are actually equally or less likely to receive this item as the AOR than teachers receiving a 2 or 3. Why is this the case? Teachers who receive a 1 in *Problem-Solving* also tend to have very low performance in other areas. In this case, the AOR choice may be driven by the rater's judgment about which areas are most important for improving the overall quality of instruction. For example, in the roughly 4,000 observations where a teacher received a 1 in *Problem-Solving*, only 7.5% of the time was *Problem-Solving* the chosen AOR. Instead, the AOR was much more often *Standards and Objectives* (18%), *Presenting Instructional Content* (17%), *Activities and Materials* (11%), or *Lesson Structure and Pacing* (10%). Raters likely view these items as representing skills that are more fundamental, those for which they can provide better coaching/support, or those that are more important for improving student achievement.

Figure 6 also helps to allay concerns that the novice–veteran AOR gaps reflect raters' pre-existing beliefs that, for instance, novice teachers should focus on improving classroom management skills. If this were the key driver of the patterns in Figure 2, we should observe large differences in the predicted probabilities for novices and veterans, conditional on the item score received. Figure 6 shows that this is not the case—the patterns are quite similar for novices and veterans. We also conduct a more specific test of AOR assignment disparities in Appendix Table A4, which estimates the difference (by teacher experience) in the predicted probability of AOR choice for each item conditional on the full set of item scores. Across all items, the estimated gaps are very small in magnitude, again suggesting that the observed AOR disparities reflect actual differences in relative skill or performance, not raters' pre-existing beliefs.[3]

In sum, Figures 5 and 6 indicate a clear relationship between a teacher's identified AOR and their performance measured by item-level scores but also signal the raters judgement of where the teacher should most productively focus their efforts to improve. This increases our confidence that the area of refinement provides a useful signal of raters' judgements regarding specific skills on which teachers should focus.

Discussion

Despite a robust body of evidence demonstrating sharp average improvement among novice teachers in their first years on the job, we know little about the nature of this skill development. What skills are novice teachers building? Are certain teaching skills more foundational than others? Successfully answering these questions can help to inform a wide range of policies that aim to increase teacher quality, including teacher preparation,

[3] As a related check, Appendix Table A5 shows that item scores have substantially greater predictive power than teacher experience for the AOR and, more generally, teacher experience on its own is a weak predictor of the AOR.

induction, and professional development.

To our knowledge, this is the first analysis to employ large-scale quantitative data to isolate specific skills associated with the overall development of novice teachers. Using data on roughly 25,000 novice teachers, we leverage a unique feature of the Tennessee teacher observation protocol that requires school administrators to identify one of the 19 observed skills as the “area of refinement”—the skill on which the teacher should focus their improvement efforts following each observation. The AOR addresses a measurement issue common in teacher observations in which most items or domains of the rubric are highly correlated, providing little insight to teacher skill improvement. Indeed, whereas item-level scores suggest large (and roughly equal) improvement on all teaching skills, examining the AOR suggests instead that certain skills—classroom management and presenting content—drive the returns to experience for novice teachers. As demonstrated in prior work (that we also replicate here), early career improvement is largest among the initially lowest-performing teachers. We find that among novice teachers, these initially lowest-performing teachers have an AOR that is disproportionately classroom management or presenting content. These teachers experience some of the greatest improvement in these skills and in overall classroom observation scores. This concordance between the AOR and generalized effectiveness measures (i.e., average observation scores and value-added) suggests wide variation in novice teachers’ skills that translates to large effectiveness differences and, ultimately, differences in student outcomes. While low-performing novice teachers who remain in teaching improve substantially over time in these areas and in their overall effectiveness, they exit the profession at high rates. Despite large improvements among those who remain, they typically fail to completely catch up to their initially higher-performing peers. While it is possible that principals or other observers identify the AOR in ways that are inconsistent with the state guidance as the skill on which teachers should focus their improvement, we find little support for any of these alternative explanations.

Perhaps the clearest implication of these findings is to underscore the importance of classroom management skills. Teachers, principals, and teaching experts have long agreed that establishing an organized classroom where all students are engaged is fundamental to learning. These views are largely based on a conceptual understanding of teacher development, anecdotal observations, and small-scale studies, but we are unaware of any systematic evidence that links classroom management to the development of novice teachers. Our results suggest that an increased focus on classroom management skills in teacher preparation—which is by no means a novel proposal (e.g., see Greenberg et al., 2014; Kwok, 2021; Levine, 2006; Pomerance & Walsh, 2020; Robertson, 2006; Stevenson et al., 2020)—is warranted. Simply put, many novice teachers enter the profession without adequate classroom management skills. While they appear to improve in this area with experience, ameliorating this initial struggle through stronger pre-service training could meaningfully raise the average quality of new teachers and help to reduce attrition.

Improvement in pre-service teacher classroom management skills could occur in several ways, but will require teacher preparation programs to invest resources in training and clinical experiences that has been absent to date. The National Council on Teacher Quality has long advocated that teacher preparation programs should employ available evidence in their training (Pomerance & Walsh, 2020). In their review of individual programs, they find that although there has been meaningful improvement, many programs still score poorly on the use of evidence in their programs. Even among programs who indicate the use of such evidence, we know little regarding the actual classroom management performance of program graduates. A growing body of research suggests that clinical practice is an important mechanism by which teaching skills can be assessed and improved. However, many programs do not provide prospective teachers with a quality clinical practice experience nor work with school districts to select an effective mentor teacher (Pomerance & Walsh, 2020).

Our results also inform long-standing debates about the nature of teacher skill development. First, we find preliminary evidence that some skills are foundational to achieving a minimum standard of effective teaching or to the development of other skills. More specifically, we observe a tradeoff in the assigned AOR whereby administrators appear to prioritize certain skills, such as presenting content and classroom management, over others, such as creating opportunities for student problem-solving and high-level thinking, even when the latter receive lower scores. One interpretation of this tradeoff is that raters view classroom management and presenting content as more fundamental or foundational skills, which should be prioritized over higher-order skills.

Relatedly, our results provide one hypothesis for the substantial variability in the returns to experience (as demonstrated in Table 1). Grouped by initial performance, the top quartile of novice teachers exhibit little to no subsequent improvement, on average. While mean reversion is one likely contributor, the AOR patterns (Figure 4) suggest that these higher-performing novice teachers enter the profession with stronger skills. For them, overall improvement requires strengthening higher-order teaching skills that are perhaps more difficult to master and for which they may receive less support.

Though our study produces new insight into teacher skill development, a number of questions remain unanswered. First, while we highlight the importance of *item salience* for the AOR, we are unable to observe exactly how administrators make this determination. Certain items, such as *Questioning*, appeared to have very high salience, which could be explained in multiple ways. Administrators may believe this skill is particularly important for effective teaching or improving student outcomes. Alternatively, *Questioning* may have high salience because performance in this area is easily observable during a classroom observation, whereas other areas are more difficult to observe (e.g., *Student Work*). Finally, administrators may prioritize skills for which they feel comfortable providing feedback to the teacher, even if improvement in other areas would be more impactful. Unfortunately, the nature of our data prevents us from understanding this process. Future studies that document how raters—particularly embedded actors like school administrators—make sense of and implement classroom observation protocols would help to pinpoint the specific drivers of the patterns we document here.

Future work should also aim to more closely examine the high item-level correlations in rubric-based classroom observation scores. To what extent does this represent the true correlation of skills? Is this pattern the byproduct of asking raters to focus on too many items at once? Such questions arise even in low-stakes settings with highly trained raters, such as the MET project (Kane & Staiger, 2012). A better understanding of this phenomenon would have important research and policy implications, including for our efforts here to understand early career skill development. Finally, our results suggest, but with a weak causal warrant, that improving classroom management skills during teacher preparation or very early in a teacher's career could catalyze the development of other skills. Using natural experiments, such as a substantial expansion of classroom management skills in some preparation programs, or a more structured field experiment has the potential to provide stronger evidence of the sequential nature of teacher skill development.

References

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

Bartanen, B., & Kwok, A. (2021). Examining Clinical Teaching Observation Scores as a Measure of Preservice Teacher Quality. *American Educational Research Journal*, *58*(5), 1–33.

Bell, C. A., James, J. K., Taylor, E. S., & Wyckoff, J. (2023). *Measuring returns to experience using supervisor ratings of observed performance: The case of classroom teachers* (tech. rep.). National Bureau of Economic Research.

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.

Cappella, E., Aber, J. L., Kim, H. Y., Gitomer, D., & Bell, C. (2016). Teaching beyond achievement tests: Perspectives from developmental and education science. *Handbook of research on teaching*, *5*, 249–347.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of education review*, *26*(6), 673–682.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management*, *34*(2), 267–297.

Dickenson, T. S., Hodges, T. E., Watts, J., Brown, A., Fan, X., Consortium, S. C. T. E. A., et al. (2022). Summary of results for the south carolina teacher exit survey from the 2021-2022 pilot administration.

Feiman-Nemser, S., Schwille, S., Carver, C., & Yusko, B. (1999). A conceptual review of literature on new teacher induction.

Ferguson, R. F., & Danielson, C. (2015). How framework for teaching and tripod 7cs evidence distinguish key components of effective teaching. *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*, 98–143.

Freeman, J., Simonsen, B., Briere, D. E., & MacSuga-Gage, A. S. (2014). Pre-service teacher training in classroom management: A review of state accreditation policy and teacher preparation programs. *Teacher Education and Special Education*, *37*(2), 106–120.

Fuller, F. F., & Bown, O. H. (1975). Becoming a teacher. *Teachers College Record*, *76*(6), 25–52.

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., Uekawa, K., Falk, A., Bloom, H. S., Doolittle, F., et al. (2008). The impact of two professional development interventions on early reading instruction and achievement. ncee 2008-4030. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., & Borman, G. D. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development. ncee 2016-4010. *National Center for Education Evaluation and Regional Assistance*.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., et al. (2011). Middle school mathematics professional development impact study: Findings after the second year of implementation. ncee 2011-4024. *National Center for Education Evaluation and Regional Assistance*.

Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? a meta-analysis of randomized field studies. *Review of Research in Education*, *43*(1), 106–137.

Gray, L., & Taie, S. (2015). Public school teacher attrition and mobility in the first five years: Results from the first through fifth waves of the 2007-08 beginning teacher longitudinal study. first look. nces 2015-337. *National center for education statistics*.

Greenberg, J., Putman, H., & Walsh, K. (2014). Training our future teachers: Classroom management. revised. *National Council on Teacher Quality*.

Grissom, J. A., & Loeb, S. (2017). Assessing principals' assessments: Subjective evaluations of teacher effectiveness in low-and high-stakes environments. *Education Finance and Policy*, *12*(3), 369–395.

Grossman, P. L. (1992). Why models matter: An alternate view on professional growth in teaching. *Review of educational research*, *62*(2), 171–179.

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.

Hammerness, K. (2011). Classroom management in the United States: A view from New York City. *Teaching Education*, *22*(2), 151–167.

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

Hunter, S. B. (2018). *History of team teacher evaluation policy* (tech. rep.). Tennessee Education Research Alliance.

Kagan, D. M. (1992). Professional growth among preservice and beginning teachers. *Review of educational research*, *62*(2), 129–169.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? evidence from new york city. *Economics of Education review*, *27*(6), 615–631.

Kane, T. J., & Staiger, D. O. (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. research paper. met project. *Bill & Melinda Gates Foundation*.

Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547–588.

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

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis*, *36*(4), 476–500.

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315–347.

Kwok, A. (2021). Managing classroom management preparation in teacher education. *Teachers and Teaching*, *27*(1-4), 206–222.

Levine, A. (2006). Educating school teachers. *Education Schools Project*.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs stem instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293.

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.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–151.

Papay, J. P., & Laski, M. (2021). *Understanding the dynamics of teacher productivity development: Evidence on teacher improvement in tennessee* (tech. rep.). Working paper.

Pomerance, L., & Walsh, K. (2020). 2020 teacher prep review: Clinical practice & classroom management. *National Council on Teacher Quality*.

Redding, C., & Nguyen, T. D. (2020). Recent trends in the characteristics of new teachers, the schools in which they teach, and their turnover rates. *Teachers College Record*, *122*(7), 1–36.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458.

Robertson, M. (2006). Why novice teachers leave. *Principal Leadership*, *6*(8), 33–36.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

Ronfeldt, M. (2021). Links among teacher preparation, retention, and teaching effectiveness. evaluating and improving teacher preparation programs. *National Academy of Education*.

Russ, R. S., Sherin, B. L., Sherin, M. G., et al. (2016). What constitutes teacher learning. *Handbook of research on teaching*, *5*, 391–438.

Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American educational research journal*, *41*(3), 681–714.

Stevenson, N. A., VanLone, J., & Barber, B. R. (2020). A commentary on the misalignment of teacher education and the need for classroom behavior management skills. *Education and Treatment of Children*, *43*(4), 393–404.

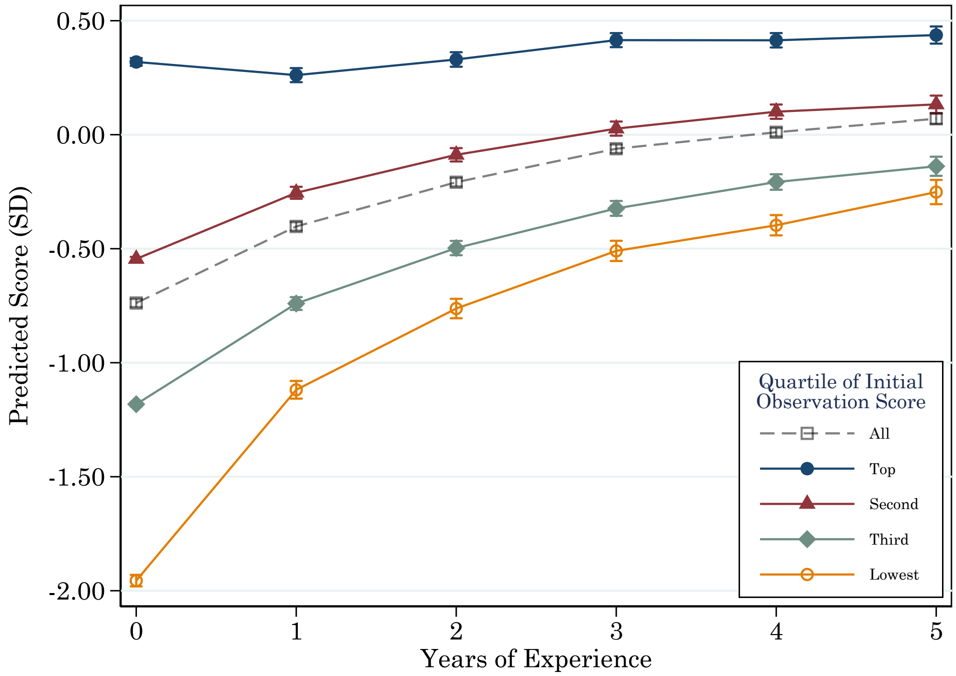
Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review*, *102*(7), 3628–3651.

Watzke, J. L. (2007). Longitudinal research on beginning teacher development: Complexity as a challenge to concerns-based stage theory. *Teaching and Teacher Education*, *23*(1), 106–122.

Wei, R. C., Darling-Hammond, L., & Adamson, F. (2010). Professional development in the united states: Trends and challenges.

Yoon, K. S., Duncan, T., Lee, S. W.-Y., Scarloss, B., & Shapley, K. L. (2007). Reviewing the evidence on how teacher professional development affects student achievement. issues & answers. rel 2007-no. 033. *Regional Educational Laboratory Southwest (NJ1)*.

(a) *Observation Scores*



(b) *Value-Added*

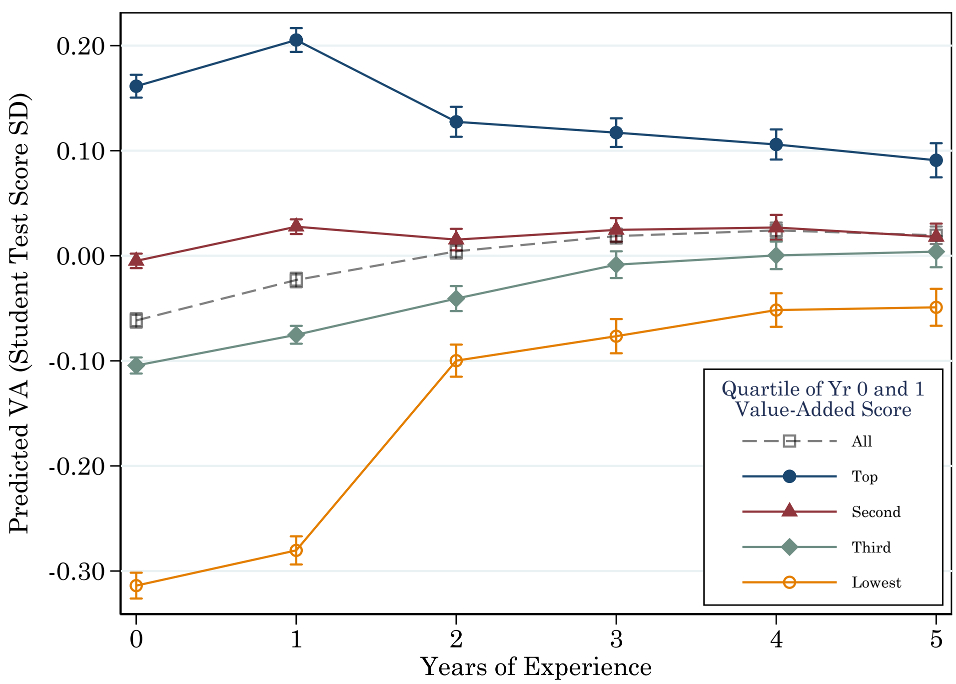
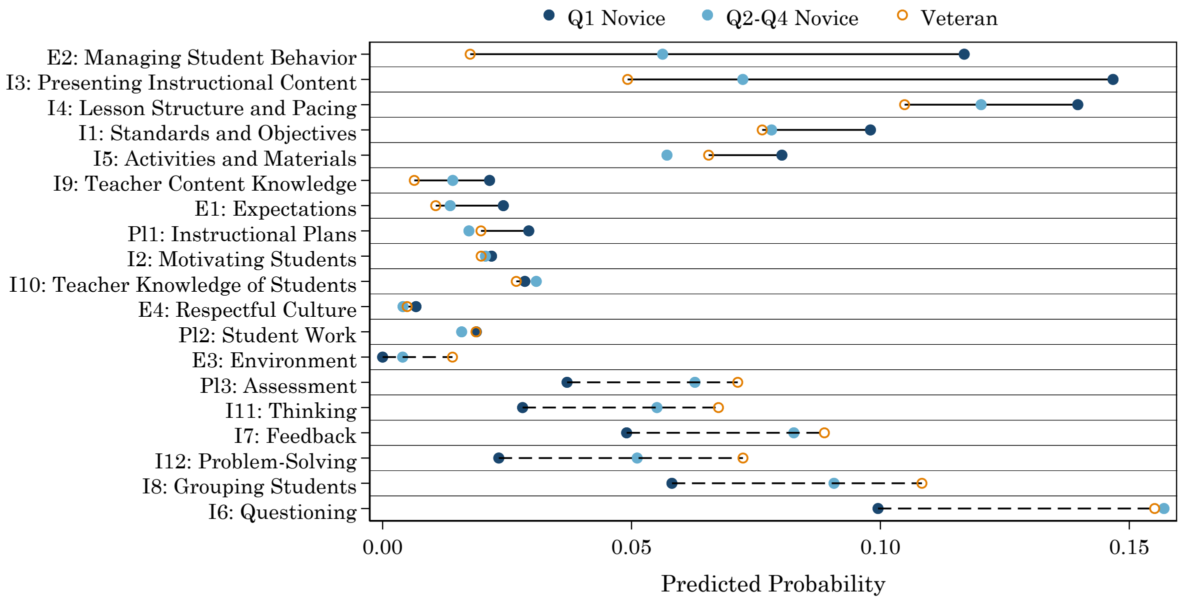


Figure 1

Returns to Experience by Initial Performance

Notes: Plots show predicted margins for the outcome variable listed in the panel header and the group defined in the legend. These predicted margins come from a least squares regression model that predicts the outcome as a function of dummy variables for each year of experience and year fixed effects. The sample is restricted to Tennessee teachers for whom we observe their first year of teaching and who remain until their sixth year (i.e., a balanced panel). Additionally, we include only the first six years of a teacher's career in the model. In Panel A, the lines are defined by performance in the teacher's first year (i.e., years of experience = 0). In Panel B, we average the first two years of value-added to reduce measurement error. Errors bars show 95% confidence intervals.

(a) *Area of Refinement*



(b) *Item Scores*

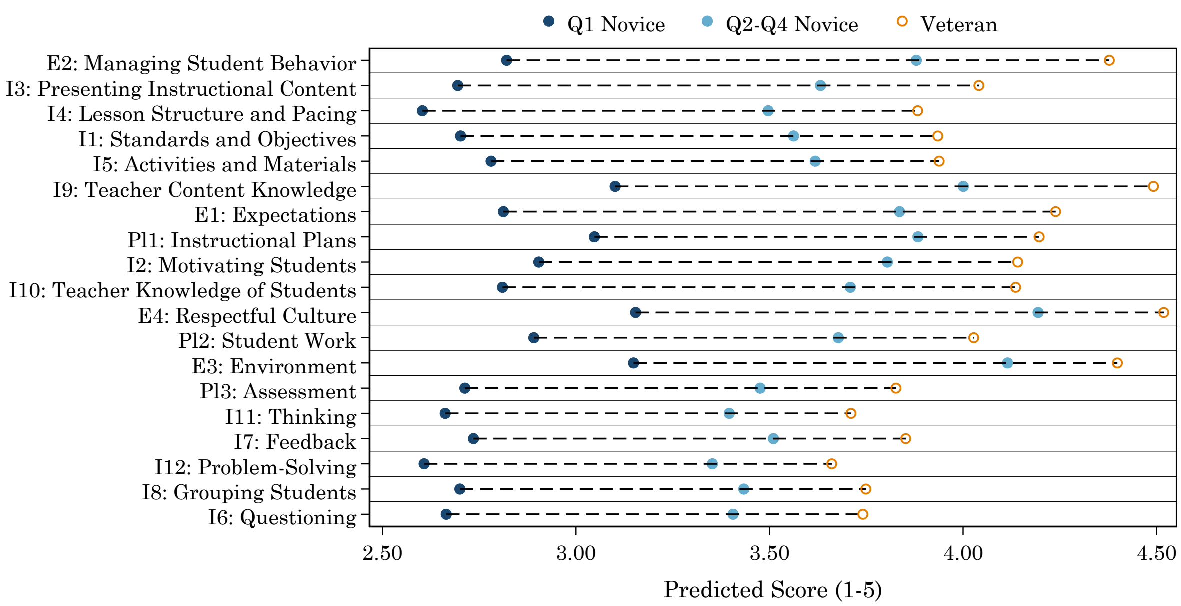


Figure 2

Comparing Novice and Veteran Teachers

Notes: Estimates shown (markers) are predicted margins for Q1 novice (first-year teachers scoring in bottom 25% of average observation scores among first-year teachers), Q2–Q4 novice, and veteran (10+ years of experience) teachers, respectively, from least squares regression models. Panel A models binary indicators for the identified area of refinement. Panel B models the 1 to 5 discrete rating for each item. Lines between markers are drawn for ease of comparison. We order the items according to the size of the Q1 novice vs. veteran difference. Solid lines indicate a positive gap (novices more likely to be flagged on the item) and dashed lines indicate a negative gap. Dependent variable is a binary indicator for whether the teacher was flagged for refinement on the listed rubric item. Unit of observation is teacher-year-instance (i.e., each individual classroom observation). Instances are weighted to sum to one within teacher-by-year-by-rubric domain (rubric domains are instruction, planning, and environment), where the within-teacher-year-domain weight is the number of rubric items that were assessed in by the rater in the observation instance divided by the total rubric items assessed over the entire school year for observation instances that included the given domain. Models include year fixed effects. See Appendix Table A2 (columns 1 and 4) for full results, including standard errors.

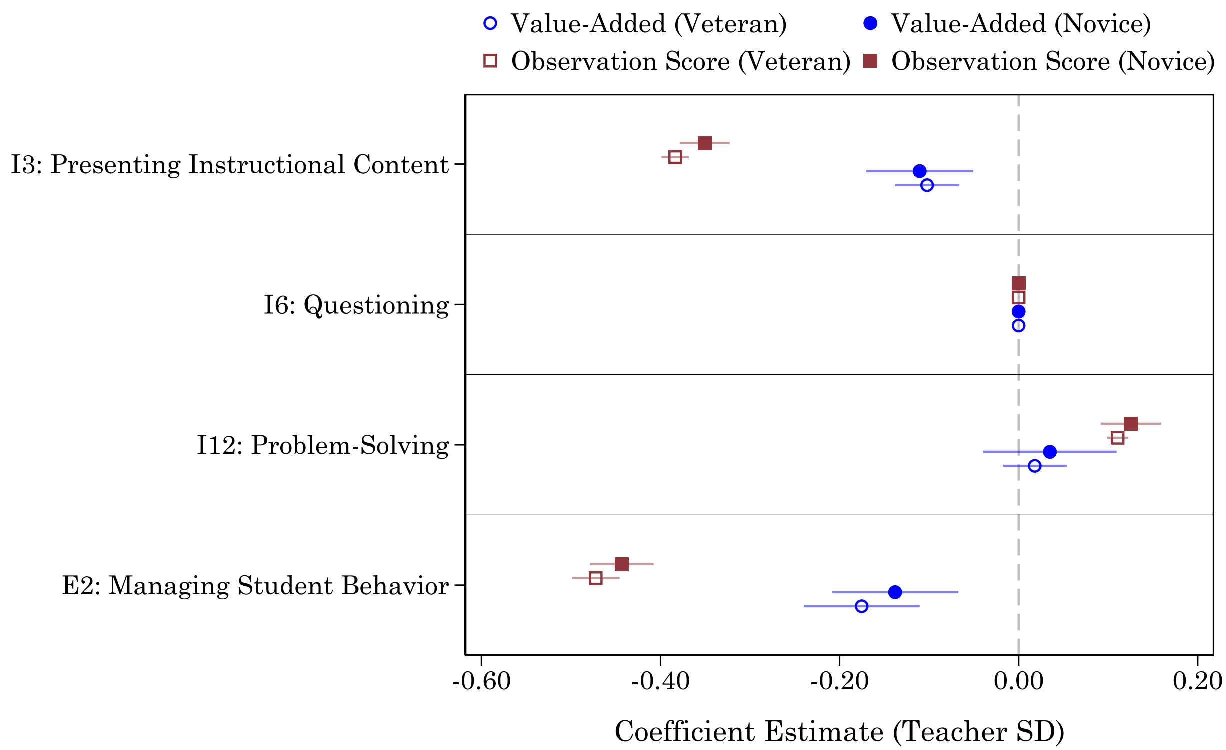


Figure 3

Area of Refinement and Teacher Effectiveness

Notes: Plot shows regression coefficients and 95% confidence intervals from models that predict value-added or average observation score as a function of the AOR for a given observation in the same year. Each model is weighted according to the procedure described in the methods section. In each model, *Questioning* is the omitted category, such that the coefficients for the 18 other items represent the predicted difference (x-axis) in effectiveness between a teacher whose AOR is the indicated item (y-axis) and a teacher whose AOR is *Questioning*. Models are estimated separately for novices and veterans. Models include fixed effects for school-by-rater-by-year and the combination of domains included in the observation. While the full set of 19 items is included in each model, we restrict the results shown to the four focal items. The full set of results is shown in Appendix Figure A3.

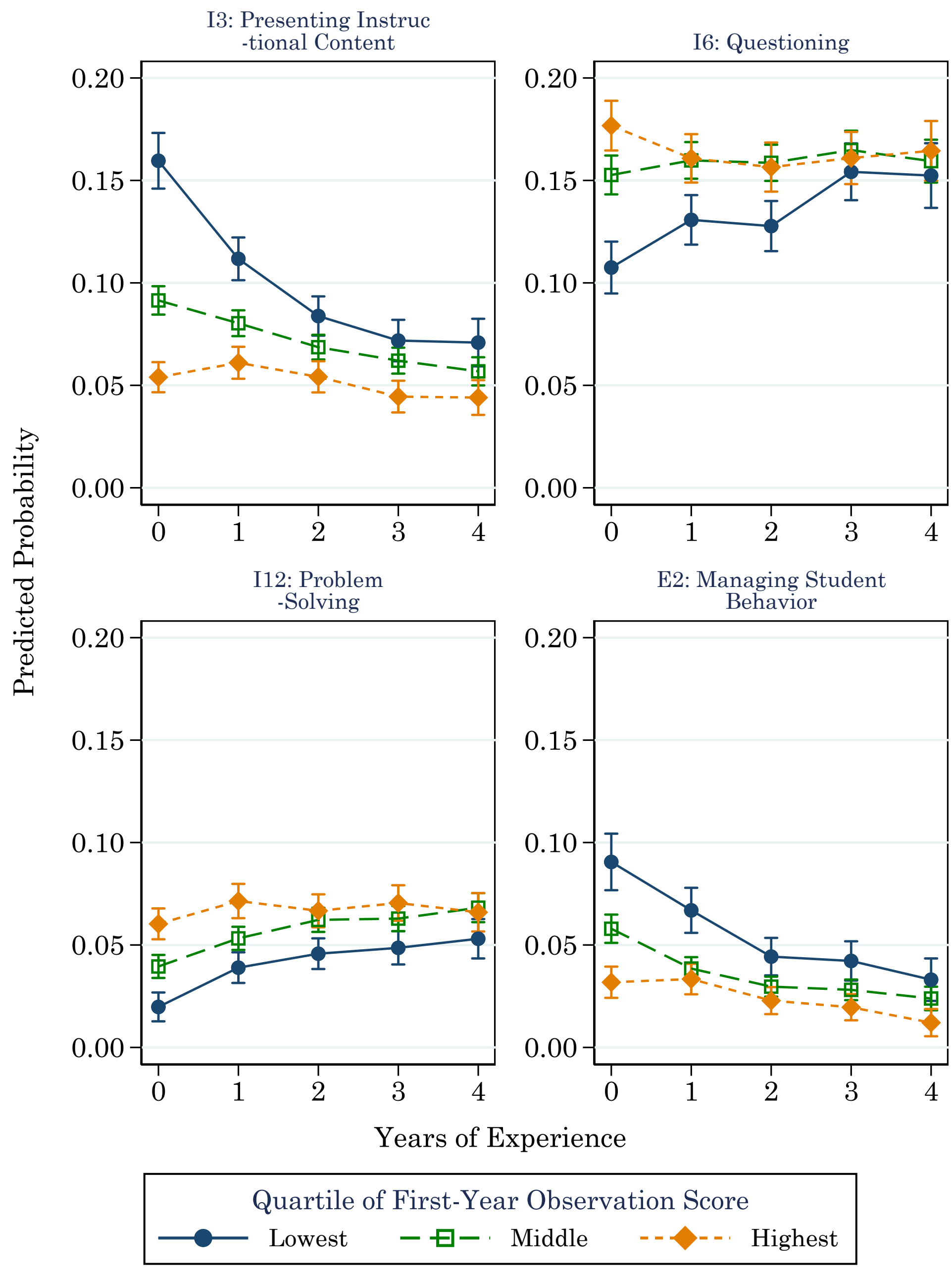


Figure 4

Area of Refinement and Teacher Experience

Notes: Plots show predicted margins from linear probability models that regress a binary indicator for whether the listed item (header) was the area of refinement for a given observation on binary indicators for teacher experience (up to 4 years) and their interaction with the first-year observation score quartile groupings (see legend). Covariates include fixed effects for year and the combination of observed rubric domains. The unit of observation is teacher-year-observation. Observations are weighted with domain-specific weights per the procedure described in the methods section, where the total weight for a teacher-by-year is 1.0. The margins represent the predicted probability that an item was chosen as the AOR in an observation where all 19 rubric items are assessed. The sample is restricted to teachers for whom we observe their first year of teaching and who remain until their fifth year. Errors bars show 95% confidence intervals.

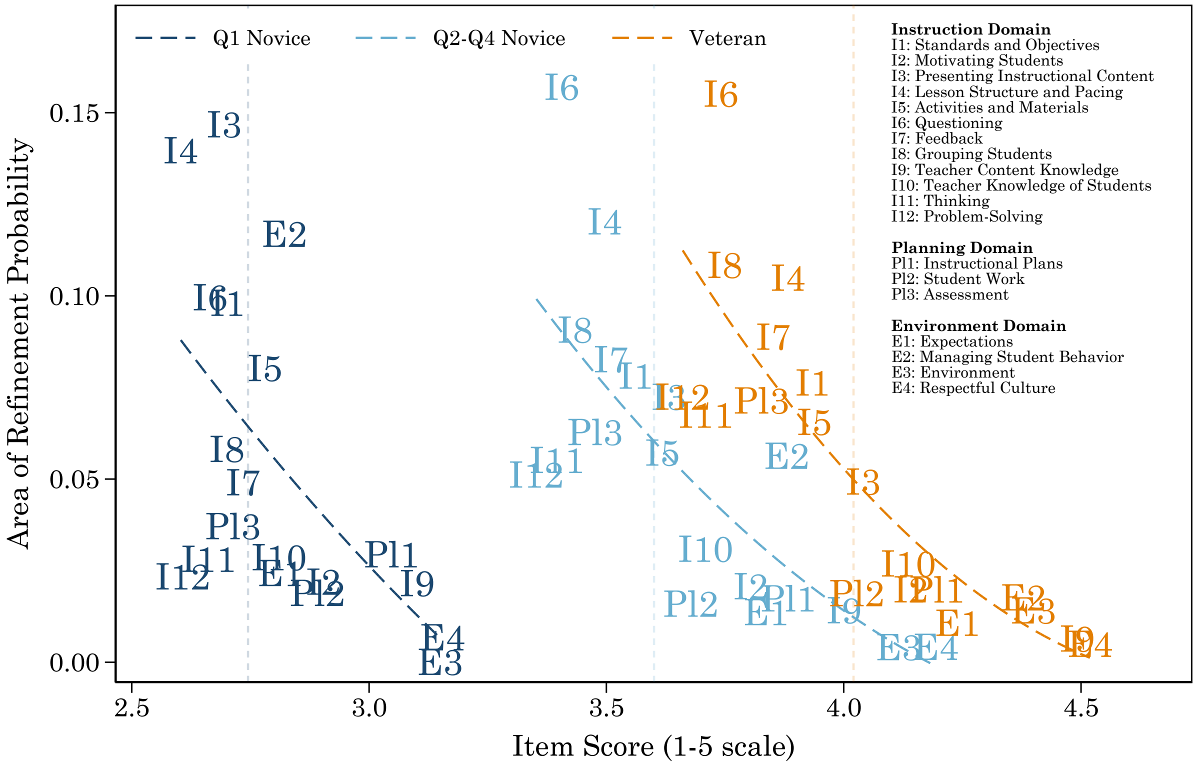


Figure 5

Comparing Area of Refinement Frequency and Item-Level Scores

Notes: Plot shows the probability that an item is chosen as the AOR in an observation where all 19 items are assessed (y-axis) versus the mean score for that item (x-axis) on the 1–5 rating scale. This relationship is plotted separately for Q1 novices, Q2–Q4 novices, and veterans. These quartiles refer to the average observation score among first-year teachers. The markers refer to each of the 19 rubric items, which are defined in the legend. The darker dashed lines show a simple quadratic fit for the 19 items. The vertical dashed lines shown the average observation score for each of the three groups.

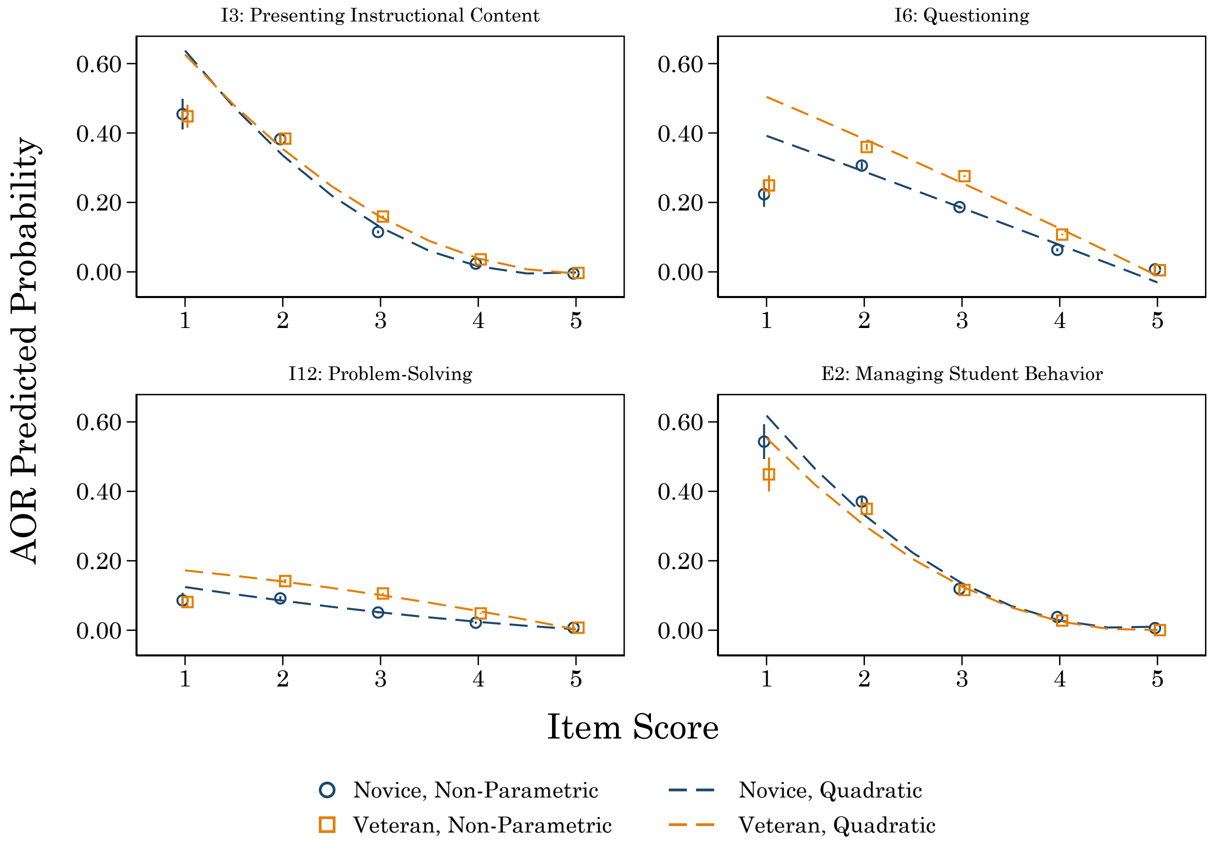


Figure 6

Predicting Area of Refinement Using Item-Level Scores From the Same Observation

Notes: Plot shows the predicted margins from linear probability models where the dependent variable is a binary indicator for whether the indicated item (header) is chosen as the area of refinement. This binary indicator is regressed on a set of binary indicators for the 1 to 5 discrete rating assigned to that same item (non-parametric results). We estimate this relationship separately for novice and veteran teachers. We also show the quadratic fit from a parallel specification that treats the item score as continuous. Models include school-rater-year fixed effects and fixed effects for the combination of rubric domains observed. The unit of observation is teacher-year-observation. Observations are weighted with domain-specific weights per the procedure described in the methods section, where the total weight for a teacher-by-year is 1.0. The margins represent the predicted probability that, in an observation where all 19 items are assessed, the listed item was chosen as the AOR as a function of the 1–5 rating given to that item. Error bars show 95% confidence intervals.

Table 1

Novice–Veteran Area of Refinement Gaps Across Different Model Specifications

	Q1 Novice–Vet Diff			Q2–Q4 Novice–Vet Diff		
	(1)	(2)	(3)	(4)	(5)	(6)
I3: Presenting Instructional Content	0.098***	0.077***	0.072***	0.023***	0.023***	0.026***
	(0.004)	(0.007)	(0.007)	(0.001)	(0.004)	(0.004)
I6: Questioning	-0.056***	-0.016*	-0.015*	0.002	0.017**	0.016**
	(0.003)	(0.009)	(0.009)	(0.002)	(0.007)	(0.007)
I12: Problem-Solving	-0.049***	-0.017***	-0.014**	-0.021***	-0.009*	-0.009**
	(0.001)	(0.006)	(0.006)	(0.001)	(0.005)	(0.005)
E2: Managing Student Behavior	0.099***	0.062***	0.059***	0.039***	0.033***	0.033***
	(0.004)	(0.007)	(0.007)	(0.002)	(0.004)	(0.004)
Year FE	✓	✓	✓	✓	✓	✓
Teacher FE		✓	✓		✓	✓
School FE			✓			✓

Notes: Each cell shows an estimate from a separate least squares regression predicting the area of refinement (labeled by row) as a function of teacher experience (entered categorically for each year up to 9 years, then a single 10+ years category). The coefficient estimate is for first-year teachers relative to teachers with 10+ years of experience. A positive coefficient indicates that novices are more likely than veterans to be flagged for refinement in the given area. Each column shows the results from a different specification, defined by the included fixed effects shown at the bottom of the table.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix A
Additional Tables and Figures (ONLINE SUPPLEMENTARY MATERIALS)

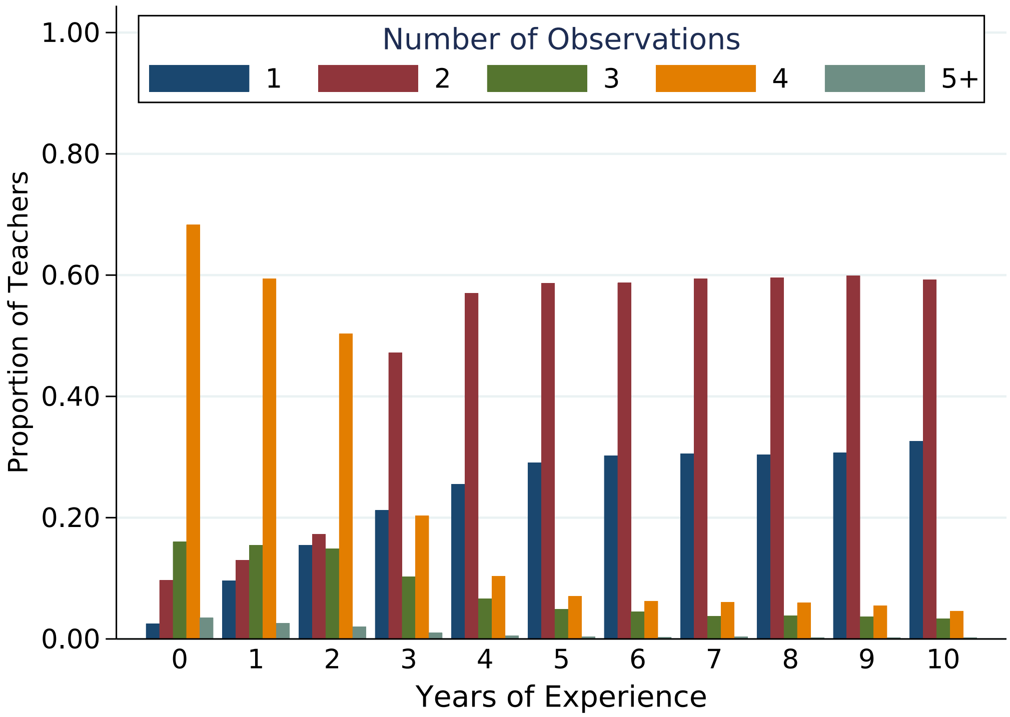
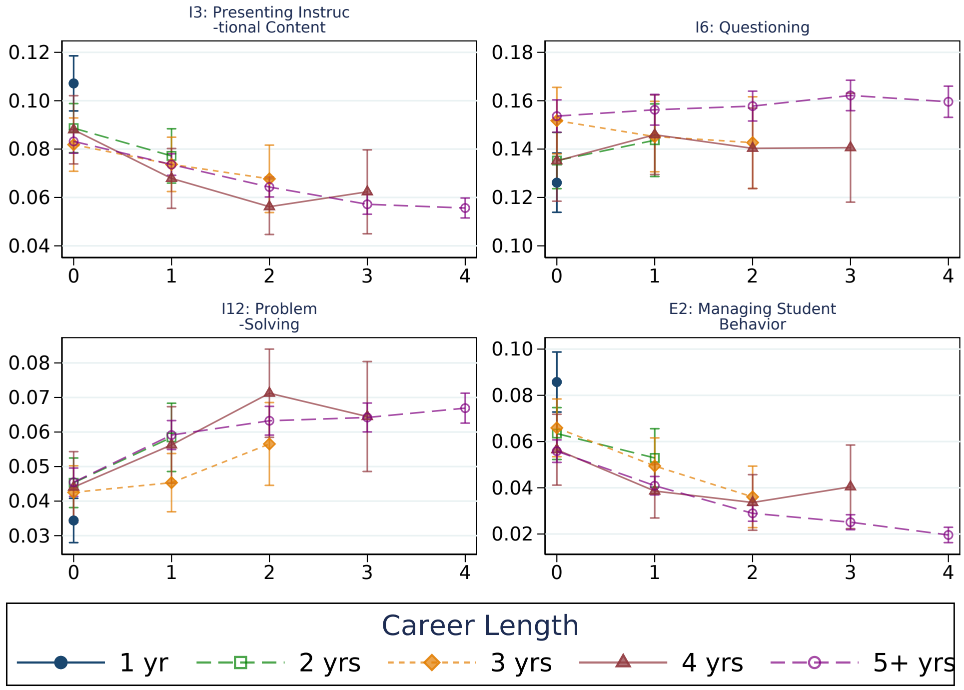


Figure A1

Number of Classroom Observations Received by Teacher Experience

Notes: Plot shows the distribution of the number of observations received by teachers with the indicated experience level (x-axis). The bars sum to 1 within each experience bin.

(a) *All Novices*



(b) *Q1 Novices*

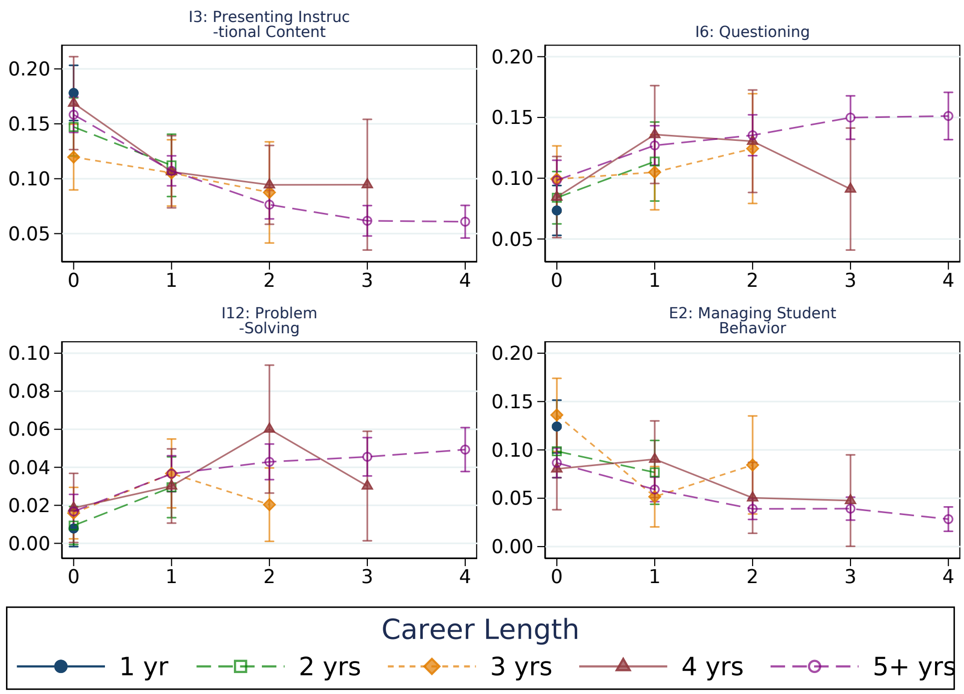


Figure A2

External Validity Check: Area of Refinement and Teacher Experience by Career Length

Notes: Plots show predicted margins from linear probability models that regress a binary indicator for whether the listed item (header) was the area of refinement for a given observation on binary indicators for teacher experience (up to 4 years) and their interaction with the career length groupings (see legend). Covariates include fixed effects for year and the combination of observed rubric domains. The unit of observation is teacher-year-observation. Observations are weighted with domain-specific weights per the procedure described in the methods section, where the total weight for a teacher-by-year is 1.0. The margins represent the predicted probability that an item was chosen as the AOR in an observation where all 19 rubric items are assessed. The sample is restricted to teachers for whom we observe their first year of teaching. Errors bars show 95% confidence intervals. The implicit "test" is to compare the slopes of the lines across groups. Differences in the levels (intercepts) are expected and captured by the teacher FE in our preferred models.

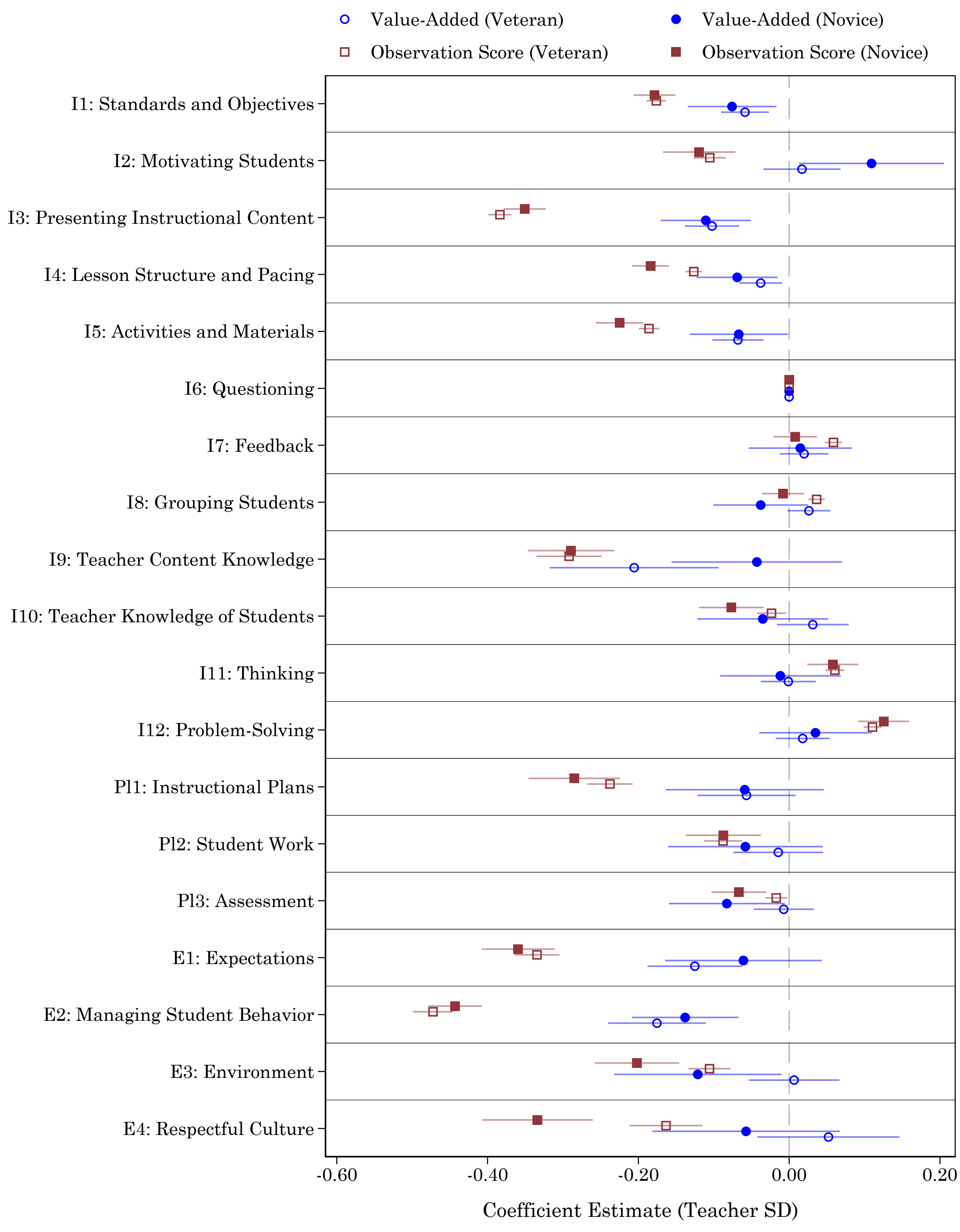


Figure A3

Area of Refinement and Teacher Effectiveness (All Items)

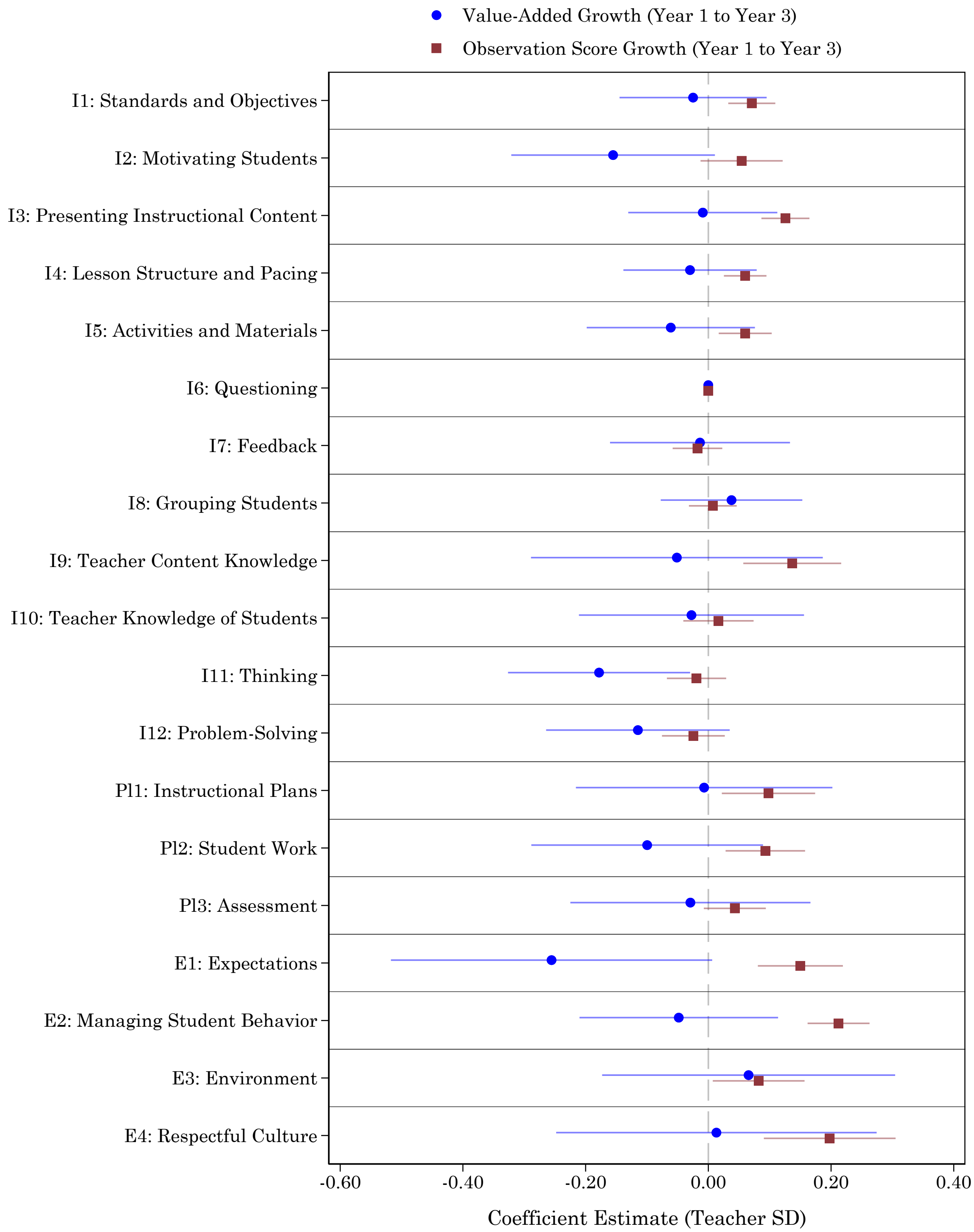


Figure A4

Area of Refinement and Growth in Teacher Effectiveness

Table A1
Descriptive Statistics

	Mean	SD	Min	Max
Teacher Characteristics				
Black	0.13			
White	0.85			
Other Race/Ethnicity	0.02			
Female	0.79			
Age	42.47	11.22	18.00	89.00
Average Observation Score (raw)	3.97	0.60	1.00	5.00
Average Observation Score (standardized)	-0.04	0.99	-5.30	1.85
Value-Added (raw)	-0.01	0.25	-2.86	2.11
Value-Added (standardized)	-0.01	1.00	-11.37	8.46
Has Value-Added Score	0.29			
Years of Experience	11.75	9.52	0.00	61.00
Experience = 0 years	0.07			
Experience = 1 years	0.06			
Experience = 2 years	0.06			
Experience = 3 years	0.05			
Experience = 4 years	0.05			
Experience = 5 years	0.04			
Experience = 6 years	0.04			
Experience = 7 years	0.04			
Experience = 8 years	0.04			
Experience = 9 years	0.04			
Experience = 10+ years	0.51			
School Characteristics				
Enrollment (100s)	8.44	4.87	0.01	55.20
Prop. Black Students	0.23	0.27	0.00	1.00
Prop. Hispanic Students	0.09	0.11	0.00	0.76
Prop. Gifted Students	0.02	0.03	0.00	0.56
Prop. SPED Students	0.15	0.06	0.00	1.00
Prop. FRPL Students	0.51	0.25	0.00	1.00
Elementary School	0.52			
Middle School	0.18			
High School	0.26			
Other School	0.04			
Urban School	0.31			
Suburban School	0.21			
Town School	0.17			
Rural School	0.31			

Notes: Table shows descriptive statistics for the analytic sample of teachers, defined as those with non-missing observation scores in the 2012–13 through 2018–19 school years. $N = 426,401$

Table A2

Novice–Veteran Area of Refinement Gaps Across Different Model Specifications (Full Results)

	Q1 Novice–Vet Diff			Q2–Q4 Novice–Vet Diff		
	(1)	(2)	(3)	(4)	(5)	(6)
I1: Standards and Objectives	0.022***	0.024***	0.026***	0.002	0.003	0.004
	(0.003)	(0.007)	(0.008)	(0.001)	(0.005)	(0.005)
I2: Motivating Students	0.002	0.003	0.004	0.001	0.001	-0.000
	(0.002)	(0.004)	(0.004)	(0.001)	(0.003)	(0.003)
I3: Presenting Instructional Content	0.098***	0.077***	0.072***	0.023***	0.023***	0.026***
	(0.004)	(0.007)	(0.007)	(0.001)	(0.004)	(0.004)
I4: Lesson Structure and Pacing	0.035***	-0.003	-0.001	0.015***	0.004	0.005
	(0.004)	(0.009)	(0.009)	(0.002)	(0.006)	(0.006)
I5: Activities and Materials	0.015***	0.014**	0.014**	-0.008***	-0.011**	-0.011**
	(0.003)	(0.007)	(0.007)	(0.001)	(0.005)	(0.005)
I6: Questioning	-0.056***	-0.016*	-0.015*	0.002	0.017**	0.016**
	(0.003)	(0.009)	(0.009)	(0.002)	(0.007)	(0.007)
I7: Feedback	-0.040***	-0.031***	-0.033***	-0.006***	-0.016***	-0.017***
	(0.002)	(0.007)	(0.007)	(0.001)	(0.005)	(0.005)
I8: Grouping Students	-0.050***	-0.046***	-0.043***	-0.018***	-0.023***	-0.025***
	(0.002)	(0.007)	(0.007)	(0.001)	(0.006)	(0.006)
I9: Teacher Content Knowledge	0.015***	0.008***	0.007**	0.008***	0.008***	0.009***
	(0.001)	(0.003)	(0.003)	(0.001)	(0.002)	(0.002)
I10: Teacher Knowledge of Students	0.002	-0.005	-0.004	0.004***	0.002	0.003
	(0.002)	(0.004)	(0.005)	(0.001)	(0.003)	(0.003)
I11: Thinking	-0.039***	-0.025***	-0.027***	-0.012***	-0.013***	-0.014***
	(0.002)	(0.006)	(0.006)	(0.001)	(0.005)	(0.005)
I12: Problem-Solving	-0.049***	-0.017***	-0.014**	-0.021***	-0.009*	-0.009**
	(0.001)	(0.006)	(0.006)	(0.001)	(0.005)	(0.005)
Pl1: Instructional Plans	0.010***	0.011**	0.010*	-0.002**	0.005	0.005
	(0.003)	(0.005)	(0.006)	(0.001)	(0.004)	(0.004)
Pl2: Student Work	0.000	0.001	0.002	-0.003***	0.003	0.003
	(0.002)	(0.006)	(0.006)	(0.001)	(0.004)	(0.004)
Pl3: Assessment	-0.034***	-0.018**	-0.018**	-0.009***	-0.003	-0.005
	(0.003)	(0.009)	(0.009)	(0.002)	(0.006)	(0.007)
E1: Expectations	0.014***	-0.008	-0.008	0.003***	-0.000	-0.000
	(0.003)	(0.005)	(0.005)	(0.001)	(0.004)	(0.004)
E2: Managing Student Behavior	0.099***	0.062***	0.059***	0.039***	0.033***	0.033***
	(0.004)	(0.007)	(0.007)	(0.002)	(0.004)	(0.004)
E3: Environment	-0.014***	-0.012***	-0.011**	-0.010***	-0.011***	-0.013***
	(0.002)	(0.005)	(0.005)	(0.001)	(0.003)	(0.003)
E4: Respectful Culture	0.002	0.005	0.004	-0.001	0.002	0.002
	(0.001)	(0.003)	(0.003)	(0.001)	(0.002)	(0.002)
Year FE	✓	✓	✓	✓	✓	✓
Teacher FE		✓	✓		✓	✓
School FE			✓			✓

Notes: Each cell shows an estimate from a separate least squares regression predicting the area of refinement (labeled by row) as a function of teacher experience (entered categorically for each year up to 9 years, then a single 10+ years category). The coefficient estimate is for first-year teachers relative to teachers with 10+ years of experience. A positive coefficient indicates that novices are more likely than veterans to be flagged for refinement in the given area. Each column shows the results from a different specification, defined by the included fixed effects shown at the bottom of the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3

Predicting Teacher Turnover Using Refinement Indicators

	Novices				Veterans			
	All	Exit	Xfer-W	Xfer-A	All	Exit	Xfer-W	Xfer-A
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
I1: Standards and Objectives	0.023***	0.022***	0.001	0.008	0.012***	0.006***	0.003**	0.002
	(0.008)	(0.007)	(0.006)	(0.006)	(0.003)	(0.002)	(0.002)	(0.001)
I2: Motivating Students	0.023*	0.023**	0.012	-0.005	0.015***	0.008**	0.005*	0.000
	(0.014)	(0.011)	(0.010)	(0.009)	(0.005)	(0.004)	(0.003)	(0.002)
I3: Presenting Instructional Content	0.046***	0.029***	0.011**	0.025***	0.008**	0.007***	0.002	0.004***
	(0.008)	(0.007)	(0.005)	(0.006)	(0.003)	(0.002)	(0.002)	(0.001)
I4: Lesson Structure and Pacing	0.019***	0.019***	0.005	0.005	0.003	0.003*	-0.001	0.002**
	(0.007)	(0.006)	(0.005)	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)
I5: Activities and Materials	0.021**	0.017**	-0.000	0.013**	0.007**	0.007***	-0.001	0.003**
	(0.009)	(0.007)	(0.006)	(0.006)	(0.003)	(0.002)	(0.002)	(0.001)
I6: Questioning (base)								
I7: Feedback	0.003	0.012*	-0.005	-0.003	-0.008***	-0.002	-0.004***	-0.003**
	(0.008)	(0.007)	(0.006)	(0.006)	(0.003)	(0.002)	(0.002)	(0.001)
I8: Grouping Students	-0.019**	-0.002	-0.012**	-0.010*	0.001	0.001	0.000	-0.001
	(0.008)	(0.006)	(0.005)	(0.005)	(0.003)	(0.002)	(0.002)	(0.001)
I9: Teacher Content Knowledge	0.029*	0.021	-0.010	0.018	0.025***	0.003	0.015***	0.009**
	(0.017)	(0.014)	(0.011)	(0.011)	(0.008)	(0.006)	(0.005)	(0.004)
I10: Teacher Knowledge of Students	0.009	0.021**	0.001	-0.002	-0.001	0.002	-0.001	-0.003*
	(0.012)	(0.010)	(0.009)	(0.008)	(0.004)	(0.003)	(0.002)	(0.002)
I11: Thinking	-0.016	-0.014**	-0.002	-0.003	0.001	-0.001	-0.002	0.000
	(0.010)	(0.007)	(0.007)	(0.007)	(0.003)	(0.002)	(0.002)	(0.001)
I12: Problem-Solving	-0.002	-0.002	0.003	-0.006	-0.002	-0.003	-0.002	0.000
	(0.010)	(0.008)	(0.007)	(0.007)	(0.003)	(0.002)	(0.002)	(0.001)
Pl1: Instructional Plans	0.057***	0.047***	0.012	0.021*	0.020***	0.013***	0.005	0.004*
	(0.017)	(0.015)	(0.012)	(0.011)	(0.006)	(0.005)	(0.003)	(0.002)
Pl2: Student Work	-0.009	0.002	-0.012	-0.001	0.005	0.009**	-0.003	0.002
	(0.015)	(0.012)	(0.009)	(0.010)	(0.006)	(0.004)	(0.003)	(0.002)
Pl3: Assessment	0.012	0.017*	-0.002	0.003	0.004	0.001	0.002	-0.000
	(0.011)	(0.009)	(0.007)	(0.007)	(0.003)	(0.002)	(0.002)	(0.001)
E1: Expectations	0.035***	0.043***	0.003	0.009	0.017***	0.009**	0.006*	0.005**
	(0.014)	(0.012)	(0.009)	(0.010)	(0.006)	(0.005)	(0.004)	(0.003)
E2: Managing Student Behavior	0.065***	0.057***	0.011	0.021***	0.011**	0.004	0.003	0.008***
	(0.010)	(0.009)	(0.007)	(0.007)	(0.005)	(0.004)	(0.003)	(0.002)
E3: Environment	0.019	0.017	-0.004	0.014	0.002	-0.004	0.005	0.002
	(0.017)	(0.015)	(0.011)	(0.012)	(0.006)	(0.004)	(0.004)	(0.002)
E4: Respectful Culture	0.042*	0.021	-0.002	0.043**	0.011	0.005	0.010	0.006
	(0.022)	(0.018)	(0.016)	(0.018)	(0.010)	(0.007)	(0.007)	(0.005)
Observations	66047	56173	55022	54855	274133	255421	253464	247795
R^2	0.454	0.457	0.538	0.469	0.203	0.160	0.304	0.177

Notes: Standard errors clustered by teacher-year in parentheses. Estimated coefficients are from linear probability models predicting teacher turnover outcomes. Other control variables include school characteristics, district fixed effects, and year fixed effects. All includes all turnover types. Exits are those who leave the state data system. Xfer-W are within-district transfers. Xfer-A are across-district transfers.

* p < 0.1, ** p < 0.05, *** p < 0.01.

Table A4

Predicting Area of Refinement Conditional on Item-Level Scores

	Years of Experience									
	0	1	2	3	4	5	6	7	8	9
I1: Standards and Objectives	-0.002	-0.001	-0.002	-0.001	-0.002	-0.001	-0.001	0.001	0.000	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I2: Motivating Students	-0.001	0.000	-0.000	0.001	0.000	0.001	0.002*	-0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I3: Presenting Instructional Content	0.001	0.003**	0.003**	0.002	0.002	0.002	0.003**	-0.002	0.002	0.004**
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I4: Lesson Structure and Pacing	0.002	-0.000	-0.000	-0.001	0.001	-0.001	-0.000	0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I5: Activities and Materials	-0.007***	-0.006***	-0.005***	-0.003*	-0.004***	-0.003**	-0.004**	-0.003	-0.001	-0.004**
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I6: Questioning	0.008***	0.004*	0.006**	0.007***	0.003	-0.002	0.005**	0.006**	-0.001	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)
I7: Feedback	0.003*	0.002	0.003	0.002	-0.000	0.005**	0.004**	0.000	0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I8: Grouping Students	-0.004**	-0.005***	-0.006***	-0.003	0.001	-0.000	-0.005***	-0.000	-0.001	0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I9: Teacher Content Knowledge	0.001	-0.000	0.001	-0.000	0.000	-0.001	0.000	0.000	0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I10: Teacher Knowledge of Students	-0.001	0.001	0.002**	0.003***	-0.001	0.000	0.002	0.000	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
I11: Thinking	-0.002	-0.003**	-0.002	-0.001	0.001	0.001	-0.000	-0.001	-0.002	-0.001
	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
I12: Problem-Solving	-0.003**	-0.000	0.000	-0.001	-0.000	0.000	-0.001	-0.002	-0.002	0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Pl1: Instructional Plans	-0.001	-0.001	0.000	-0.002	-0.001	0.001	-0.001	0.000	0.003**	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Pl2: Student Work	-0.000	0.002	0.000	-0.002	-0.001	-0.001	-0.002	-0.002	-0.002	0.000
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Pl3: Assessment	0.001	0.001	0.002	-0.002	-0.002	-0.002	-0.004*	-0.001	0.000	-0.004*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
E1: Expectations	-0.002	-0.001	-0.001	-0.000	0.002*	0.001	0.001	-0.001	-0.002*	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E2: Managing Student Behavior	0.008***	0.008***	0.003**	0.002	-0.001	0.002	0.002	0.002	0.003**	-0.000
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
E3: Environment	-0.003**	-0.001	-0.001	-0.001	0.001	-0.001	-0.001	0.001	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
E4: Respectful Culture	-0.002*	-0.000	0.001	0.001	-0.000	-0.001*	0.000	0.001	0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

Notes: Each row is a separate regression predicting a binary indicator for the listed area of refinement. Columns show the coefficient estimates for years of experience, relative to the omitted category of 10+ years. Models include the full set of item-level scores (entered categorically for each discrete level, 1 to 5 or a separate category for missing) across all 19 rubric items. Heteroskedasticity-robust standard errors shown in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A5

Comparing the AOR Predictive Power of Item Scores and Teacher Experience

	(1)	(2)	(3)
	Within R^2		
I1: Standards and Objectives	0.1136	0.0001	0.1159
I2: Motivating Students	0.0287	0.0000	0.0295
I3: Presenting Instructional Content	0.1148	0.0013	0.1157
I4: Lesson Structure and Pacing	0.1067	0.0003	0.1088
I5: Activities and Materials	0.0673	0.0001	0.0697
I6: Questioning	0.0669	0.0001	0.0698
I7: Feedback	0.0337	0.0001	0.0358
I8: Grouping Students	0.0556	0.0005	0.0595
I9: Teacher Content Knowledge	0.0316	0.0006	0.0314
I10: Teacher Knowledge of Students	0.0276	0.0001	0.0285
I11: Thinking	0.0206	0.0004	0.0226
I12: Problem-Solving	0.0168	0.0006	0.0190
Pl1: Instructional Plans	0.0435	0.0001	0.0434
Pl2: Student Work	0.0123	0.0002	0.0124
Pl3: Assessment	0.0264	0.0003	0.0267
E1: Expectations	0.0178	0.0011	0.0180
E2: Managing Student Behavior	0.1233	0.0062	0.1237
E3: Environment	0.0217	0.0000	0.0221
E4: Respectful Culture	0.0150	0.0002	0.0152
Item Score	✓		✓
Experience		✓	✓

Notes: Each cell is a separate regression that predicts via linear probability model a binary indicator for whether the indicated item (row) is the AOR as a function of the item score (column 1), teacher experience (column 2), or both (column 3). The estimates shown are the within R-squared, which is the proportion of within-school-rater-year variation explained by the model.

Appendix B
Details on Weighting Scheme (ONLINE SUPPLEMENTARY MATERIALS)

Figures B1 (novice teachers) and B2 (veteran teachers) show estimated frequencies of refinement indicators with and without applying the preferred weighting scheme. We also show tabulations from a "simple" weighting scheme that simply sets the weight equal to one if the relevant domain area was assessed in the given observation, and zero otherwise. Intuitively, the unweighted tabulations simply reflect the frequency of the refinement over *all* observations, regardless of whether the relevant domain area (e.g., planning for the assessment item) was even assessed. These probabilities will sum to 100% across all items. Applying simple weights increases the frequency of all items by the same amount within domains (in percentage terms), because this is a simple reduction in the denominator. Preferred weights make additional adjustments for the number of total items that were observed in a particular observation. When a particular item was chosen for the area of refinement over more alternatives, it receives greater weight.

Figure B1 illustrates the logic of the weighting. Under simple weights, all frequencies increase, but the adjustment is larger for planning and environment items because they are more likely than instruction items to not be rated on an observation (observations are typically instruction and either planning or environment). Preferred weights adjust for the fact that some observations assess far fewer items, which inflates the observed frequency of certain items. Here, the adjustments only substantively affect planning and environment because these domains have fewer items (three and four, respectively). While not particularly common, some novice teachers have observations where *only* planning or environment is assessed. This inflates the frequency of certain items, such as assessment. The large difference between the probability of assessment under simple and preferred weighting reflects that assessment is often chosen as the area of refinement in "planning-only" observations. In fact, this disparity is even greater for Pl2: Student Work, where preferred weighting reduces the probability to roughly the weighted level. This demonstrates that Student Work is rarely chosen as an area of refinement in observations that include instruction and/or environment items. By contrast, Managing Student Behavior remains fairly frequent under preferred weighting because it is more often observed (relative to other items, such as Expectations) in observations that include instruction and/or planning items.

For veteran teachers in Figure B2, there are few substantive differences between frequencies under simple or preferred weighting. This is because observations that include planning or environment as the only assessed domain are very rare. Nearly all observations include at least two domains, which results in a similar number of possible items (15, 16, or 19), and thus similar weights under simple and preferred schemes.

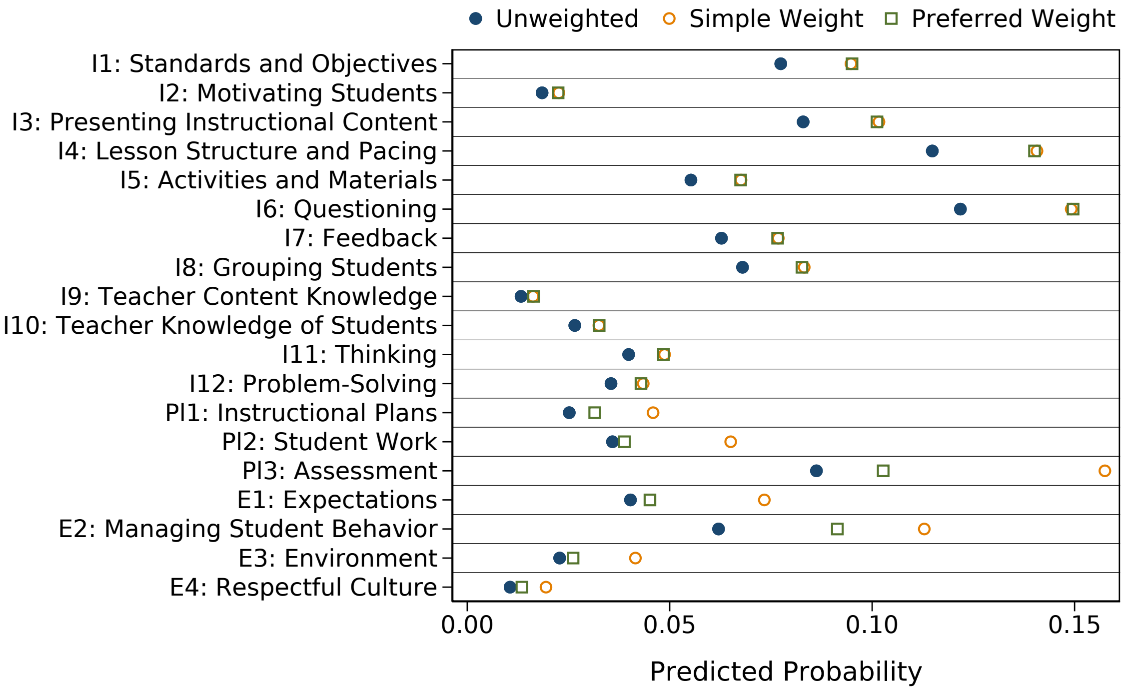


Figure B1

Frequency of Refinement Areas, Unweighted vs. Weighted (Novices)

Notes: Sample includes first-year teachers. Estimates show the frequency of areas of refinement under different weighting schemes. Simple weighting sets $weight = 0$ observations where the relevant rubric domain for an item (e.g., Instruction domain for I6: Questioning) was not assessed, and $weight = 1$ otherwise. Preferred weighting follows the procedure described in the main text.

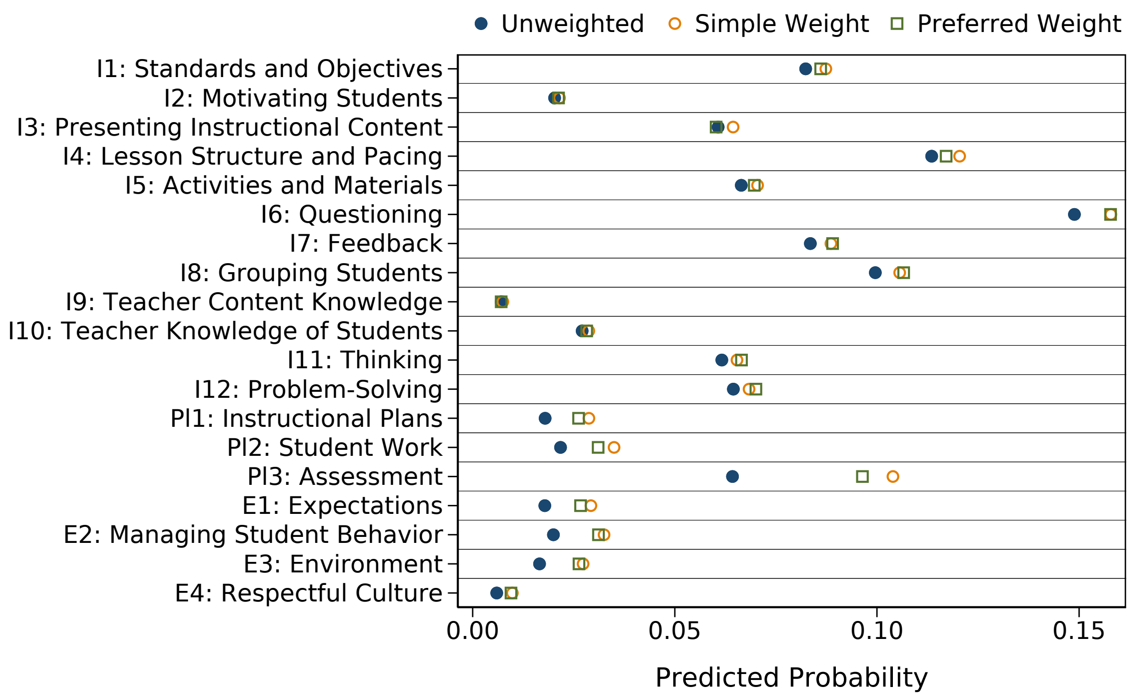


Figure B2

Frequency of Refinement Areas, Unweighted vs. Weighted (Veterans)

Notes: Sample includes teachers with 10 or more years of experience. Estimates show the frequency of areas of refinement under different weighting schemes. Simple weighting sets *weight* = 0 observations where the relevant rubric domain for an item (e.g., Instruction domain for I6: Questioning) was not assessed, and *weight* = 1 otherwise. Preferred weighting follows the procedure described in the main text.

Appendix C
TEAM Rubric (ONLINE SUPPLEMENTARY MATERIALS)

General Educator Rubric: Instruction

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Standards and Objectives	• All learning objectives are clearly and explicitly communicated, connected to state standards and referenced throughout lesson. • Sub-objectives are aligned and logically sequenced to the lesson's major objective. • Learning objectives are: (a) consistently connected to what students have previously learned, (b) know from life experiences, and (c) integrated with other disciplines. • Expectations for student performance are clear, demanding, and high. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Most learning objectives are communicated, connected to state standards and referenced throughout lesson. • Sub-objectives are mostly aligned to the lesson's major objective. • Learning objectives are connected to what students have previously learned. • Expectations for student performance are clear. • There is evidence that most students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.	• Few learning objectives are communicated, connected to state standards and referenced throughout lesson. • Sub-objectives are inconsistently aligned to the lesson's major objective. • Learning objectives are rarely connected to what students have previously learned. • Expectations for student performance are vague. • There is evidence that few students demonstrate mastery of the daily objective that supports significant progress towards mastery of a standard.
Motivating Students	• The teacher consistently organizes the content so that it is personally meaningful and relevant to students. • The teacher consistently develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher regularly reinforces and rewards effort.	• The teacher sometimes organizes the content so that it is personally meaningful and relevant to students. • The teacher sometimes develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher sometimes reinforces and rewards effort.	• The teacher rarely organizes the content so that it is personally meaningful and relevant to students. • The teacher rarely develops learning experiences where inquiry, curiosity, and exploration are valued. • The teacher rarely reinforces and rewards effort.
Presenting Instructional Content	Presentation of content always includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • effective modeling of thinking process by the teacher and/or students guided by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; • no irrelevant, confusing, or non-essential information.	Presentation of content most of the time includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; • no irrelevant, confusing, or non-essential information.	Presentation of content rarely includes: • visuals that establish the purpose of the lesson, preview the organization of the lesson, and include internal summaries of the lesson; • examples, illustrations, analogies, and labels for new concepts and ideas; • modeling by the teacher to demonstrate performance expectations; • concise communication; • logical sequencing and segmenting; • all essential information; • no irrelevant, confusing, or non-essential information.
Lesson Structure and Pacing	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end.	• The lesson starts promptly. • The lesson's structure is coherent, with a beginning, middle, and end.	• The lesson does not start promptly. • The lesson has a structure, but may be missing closure or introductory elements.

General Educator Rubric: Instruction

	• The lesson includes time for reflection. • Pacing is brisk and provides many opportunities for individual students who progress at different learning rates. • Routines for distributing materials are seamless. • No instructional time is lost during transitions.	• Pacing is appropriate and sometimes provides opportunities for students who progress at different learning rates. • Routines for distributing materials are efficient. • Little instructional time is lost during transitions.	• Pacing is appropriate for less than half of the students and rarely provides opportunities for students who progress at different learning rates. • Routines for distributing materials are inefficient. • Considerable time is lost during transitions.
Activities and Materials	Activities and materials include all of the following: o support the lesson objectives; o are challenging; o sustain students' attention; o elicit a variety of thinking; o provide time for reflection; o are relevant to students' lives; o provide opportunities for student-to-student interaction; o induce student curiosity and suspense; o provide students with choices; o incorporate multimedia and technology; and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • In addition, sometimes activities are game-like, involve simulations, require creating products, and demand self-direction and self-monitoring. • The preponderance of activities demand complex thinking and analysis. • Texts and tasks are appropriately complex.	Activities and materials include most of the following: o support the lesson objectives; o are challenging; o sustain students' attention; o elicit a variety of thinking; o provide time for reflection; o are relevant to students' lives; o provide opportunities for student-to-student interaction; o induce student curiosity and suspense; o provide students with choices; o incorporate multimedia and technology; and o incorporate resources beyond the school curriculum texts (e.g., teacher-made materials, manipulatives, resources from museums, cultural centers, etc.). • Texts and tasks are appropriately complex.	Activities and materials include few of the following: o support the lesson objectives; o are challenging; o sustain students' attention; o elicit a variety of thinking; o provide time for reflection; o are relevant to students' lives; o provide opportunities for student to student interaction; o induce student curiosity and suspense; o provide students with choices; o incorporate multimedia and technology; and o incorporate resources beyond the school curriculum texts (e.g., teacher made materials, manipulatives, resources from museums, etc.).
Questioning	Teacher questions are varied and high-quality, providing a balanced mix of question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions require students to regularly cite evidence throughout lesson. • Questions are consistently purposeful and coherent. • A high frequency of questions is asked. • Questions are consistently sequenced with attention to the instructional goals. • Questions regularly require active responses (e.g.,	Teacher questions are varied and high-quality providing for some, but not all, question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions usually require students to cite evidence • Questions are usually purposeful and coherent. • A moderate frequency of questions asked. • Questions are sometimes sequenced with attention to the instructional goals. • Questions sometimes require active responses (e.g., whole class signaling, choral responses, or	Teacher questions are inconsistent in quality and include few question types: o knowledge and comprehension; o application and analysis; and o creation and evaluation. • Questions are random and lack coherence. • A low frequency of questions is asked. • Questions are rarely sequenced with attention to the instructional goals. • Questions rarely require active responses (e.g., whole class signaling, choral responses, or group and individual answers). • Wait time is inconsistently provided.

General Educator Rubric: Instruction

	whole class signaling, choral responses, written and shared responses, or group and individual answers). • Wait time (3-5 seconds) is consistently provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • Students generate questions that lead to further inquiry and self-directed learning. • Questions regularly assess and advance student understanding • When text is involved, majority of questions are text based	group and individual answers). • Wait time is sometimes provided. • The teacher calls on volunteers and non-volunteers, and a balance of students based on ability and sex. • When text is involved, majority of questions are text based	• The teacher mostly calls on volunteers and high-ability students.
Feedback	• Oral and written feedback is consistently academically focused, frequent, high-quality and references expectations • Feedback is frequently given during guided practice and homework review. • The teacher circulates to prompt student thinking, assess each student's progress, and provide individual feedback. • Feedback from students is regularly used to monitor and adjust instruction. • Teacher engages students in giving specific and high-quality feedback to one another.	• Oral and written feedback is mostly academically focused, frequent, and mostly high-quality. • Feedback is sometimes given during guided practice and homework review. • The teacher circulates during instructional activities to support engagement, and monitor student work. • Feedback from students is sometimes used to monitor and adjust instruction.	• The quality and timeliness of feedback is inconsistent. • Feedback is rarely given during guided practice and homework review. • The teacher circulates during instructional activities, but monitors mostly behavior. • Feedback from students is rarely used to monitor or adjust instruction.
Grouping Students	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) consistently maximize student understanding and learning efficiency. • All students in groups know their roles, responsibilities, and group work expectations. • All students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to best accomplish the goals of the lesson. • Instructional groups facilitate opportunities for students to set goals, reflect on, and evaluate their learning.	• The instructional grouping arrangements (either whole class, small groups, pairs, individual; heterogeneous or homogenous ability) adequately enhance student understanding and learning efficiency. • Most students in groups know their roles, responsibilities, and group work expectations. • Most students participating in groups are held accountable for group work and individual work. • Instructional group composition is varied (e.g., race, gender, ability, and age) to most of the time, accomplish the goals of the lesson.	• The instructional grouping arrangements (either whole-class, small groups, pairs, individual; heterogeneous or homogenous ability) inhibit student understanding and learning efficiency. • Few students in groups know their roles, responsibilities, and group work expectations. • Few students participating in groups are held accountable for group work and individual work. • Instructional group composition remains unchanged irrespective of the learning and instructional goals of a lesson.

General Educator Rubric: Instruction

Teacher Content Knowledge	• Teacher displays extensive content knowledge of all the subjects she or he teaches. • Teacher regularly implements a variety of subject-specific instructional strategies to enhance student content knowledge. • The teacher regularly highlights key concepts and ideas and uses them as bases to connect other powerful ideas. • Limited content is taught in sufficient depth to allow for the development of understanding.	• Teacher displays accurate content knowledge of all the subjects he or she teaches. • Teacher sometimes implements subject-specific instructional strategies to enhance student content knowledge. • The teacher sometimes highlights key concepts and ideas and uses them as bases to connect other powerful ideas.	• Teacher displays under-developed content knowledge in several subject areas. • Teacher rarely implements subject-specific instructional strategies to enhance student content knowledge. • Teacher does not understand key concepts and ideas in the discipline and therefore presents content in an unconnected way.
Teacher Knowledge of Students	• Teacher practices display understanding of each student's anticipated learning difficulties. • Teacher practices regularly incorporate student interests and cultural heritage. • Teacher regularly provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices display understanding of some student anticipated learning difficulties. • Teacher practices sometimes incorporate student interests and cultural heritage. • Teacher sometimes provides differentiated instructional methods and content to ensure children have the opportunity to master what is being taught.	• Teacher practices demonstrate minimal knowledge of students anticipated learning difficulties. • Teacher practices rarely incorporate student interests or cultural heritage. • Teacher practices demonstrate little differentiation of instructional methods or content.
Thinking	The teacher thoroughly teaches two or more types of thinking: o analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; o practical thinking, where students use, apply, and implement what they learn in real-life scenarios; o creative thinking, where students create, design, imagine, and suppose; and o research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. The teacher provides opportunities where students: o generate a variety of ideas and alternatives; o analyze problems from multiple perspectives and viewpoints; and o monitor their thinking to insure that they understand what they are learning, are attending to critical information, and are aware of the learning strategies that they are using and why.	The teacher thoroughly teaches one type of thinking: o analytical thinking, where students analyze, compare and contrast, and evaluate and explain information; o practical thinking, where students use, apply, and implement what they learn in real-life scenarios; o creative thinking, where students create, design, imagine, and suppose; and o research-based thinking, where students explore and review a variety of ideas, models, and solutions to problems. The teacher provides opportunities where students: o generate a variety of ideas and alternatives; and o analyze problems from multiple perspectives and viewpoints.	The teacher implements no learning experiences that thoroughly teach any type of thinking. The teacher provides no opportunities where students: o generate a variety of ideas and alternatives; or o analyze problems from multiple perspectives and viewpoints.

General Educator Rubric: Instruction

Problem-Solving	The teacher implements activities that teach and reinforce three or more of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solutions • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements activities that teach two of the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing	The teacher implements no activities that teach the following problem-solving types: • Abstraction • Categorization • Drawing Conclusions/Justifying Solution • Predicting Outcomes • Observing and Experimenting • Improving Solutions • Identifying Relevant/Irrelevant Information • Generating Ideas • Creating and Designing

General Educator Rubric: Planning

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)

Instructional Plans ☐	Instructional plans include:
	• measurable and explicit goals aligned to state content standards;
	• activities, materials, and assessments that:
	  o are aligned to state standards.
	  o are sequenced from basic to complex.
	  o build on prior student knowledge, are relevant to students' lives, and integrate other disciplines.
	  o provide appropriate time for student work, student reflection, and lesson unit and closure;
	• evidence that plan is appropriate for the age, knowledge, and interests of all learners; and
	• evidence that the plan provides regular opportunities to accommodate individual student needs.
		Instructional plans include:
		• goals aligned to state content standards;
		• activities, materials, and assessments that:
		  o are aligned to state standards.
		  o are sequenced from basic to complex.
		  o build on prior student knowledge.
		  o provide appropriate time for student work, and lesson and unit closure;
		• evidence that plan is appropriate for the age, knowledge, and interests of most learners; and
		• evidence that the plan provides some opportunities to accommodate individual student needs.
			Instructional plans include:
			• few goals aligned to state content standards;
			• activities, materials, and assessments that:
			  o are rarely aligned to state standards.
			  o are rarely logically sequenced.
			  o rarely build on prior student knowledge.
			  o inconsistently provide time for student work, and lesson and unit closure;
			• little evidence that the plan provides some opportunities to accommodate individual student needs.

Student Work ☐	Assignments require students to: • organize, interpret, analyze, synthesize, and evaluate information rather than reproduce it; • draw conclusions, make generalizations, and produce arguments that are supported through extended writing; and • connect what they are learning to experiences, observations, feelings, or situations significant in their daily lives both inside and outside of school.	Assignments require students to: • interpret information rather than reproduce it; • draw conclusions and support them through writing; and • connect what they are learning to prior learning and some life experiences.	Assignments require students to: • mostly reproduce information; • rarely draw conclusions and support them through writing; and • rarely connect what they are learning to prior learning or life experiences.
Assessment ☐	Assessment Plans: • are aligned with state content standards; • have clear measurement criteria; • measure student performance in more than three ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require extended written tasks; • are portfolio-based with clear illustrations of student progress toward state content standards; and • include descriptions of how assessment results will be used to inform future instruction.	Assessment Plans: • are aligned with state content standards; • have measurement criteria; • measure student performance in more than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); • require written tasks; and • include performance checks throughout the school year.	Assessment Plans: • are rarely aligned with state content standards; • have ambiguous measurement criteria; • measure student performance in less than two ways (e.g., in the form of a project, experiment, presentation, essay, short answer, or multiple choice test); and • include performance checks, although the purpose of these checks is not clear.

General Educator Rubric: Environment

	Significantly Above Expectations (5)	At Expectations (3)	Significantly Below Expectations (1)
Expectations	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where all students can experience success. • Students take initiative and follow through with their own work. • Teacher optimizes instructional time, teaches more material, and demands better performance from every student.	• Teacher sets high and demanding academic expectations for every student. • Teacher encourages students to learn from mistakes. • Teacher creates learning opportunities where most students can experience success. • Students complete their work according to teacher expectations.	• Teacher expectations are not sufficiently high for every student. • Teacher creates an environment where mistakes an failure are not viewed as learning experiences. • Students demonstrate little or no pride in the quality of their work.
Managing Student Behavior	• Students are consistently well-behaved and on task. • Teacher and students establish clear rules for learning and behavior. • The teacher overlooks inconsequential behavior. • The teacher deals with students who have caused disruptions rather than the entire class. • The teacher attends to disruptions quickly and firmly.	• Students are mostly well-behaved and on task, some minor learning disruptions may occur. • Teacher establishes rules for learning and behavior. • The teacher uses some techniques, such as social approval, contingent activities, and consequences, to maintain appropriate student behavior. • The teacher overlooks some inconsequential behavior, but other times addresses it, stopping the lesson. • The teacher deals with students who have caused disruptions, yet sometimes he or she addresses the entire class.	• Students are not well-behaved and are often off task. • Teacher establishes few rules for learning and behavior. • The teacher uses few techniques to maintain appropriate student behavior. • The teacher cannot distinguish between inconsequential behavior and inappropriate behavior. • Disruptions frequently interrupt instruction.
Environment	The classroom: • welcomes all members and guests. • is organized and understandable to all students. • supplies, equipment, and resources are all easily and readily accessible. • displays student work that frequently changes. • is arranged to promote individual and group learning.	The classroom: • welcomes most members and guests. • is organized and understandable to most students. • supplies, equipment, and resources are accessible. • displays student work. • is arranged to promote individual and group learning.	The classroom: • is somewhat cold and uninviting. • is not well organized and understandable to students. • supplies, equipment, and resources are difficult to access. • does not display student work. • is not arrange to promote group learning.
Respectful Culture	• Teacher-student interactions demonstrate caring and respect for one another. • Students exhibit caring and respect for one another. • Positive relationships and interdependence characterize the classroom.	• Teacher-student interactions are generally friendly, but may reflect occasional inconsistencies, favoritism, or disregard for students' cultures. • Students exhibit respect for the teacher, and are generally polite to each other. • Teacher is sometimes receptive to the interests and opinions of students.	• Teacher-student interactions are sometimes authoritarian, negative, or inappropriate. • Students exhibit disrespect for the teacher. • Student interaction is characterized by conflict, sarcasm, or put-downs. • Teacher is not receptive to interests and opinions of students.