CALDER

NATIONAL
CENTER *for* ANALYSIS *of* LONGITUDINAL DATA *in* EDUCATION RESEARCH

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

A program of research by the American Institutes for Research with Duke University, Northwestern University, Stanford University, University of Missouri-Columbia, University of North Carolina at Chapel Hill, University of Texas at Dallas, and University of Washington

AIR

How Well Do Professional Reference Ratings Predict Teacher Performance?

Dan Goldhaber
Cyrus Grout
Malcolm Wolff

WORKING PAPER No. 272-1022 • October 2022

How Well Do Professional Reference Ratings Predict Teacher Performance?

Dan Goldhaber
American Institutes for Research/CALDER
University of Washington/CEDR

Cyrus Grout
University of Washington/CEDR

Malcolm Wolff
University of Washington

Contents

Acknowledgments

This work is supported by the Institute of Education Sciences (grant # R305A170060). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the institutions to which the authors are affiliated or the studies funder. We are grateful for comments on an earlier draft from Ben Backes, Edie Harding and Susanna Loeb. Correspondence regarding this article should be addressed to Dan Goldhaber, National Center for the Analysis of Longitudinal Data in Education Research, American Institutes for Research, 3876 Bridge Way N, Suite 201, Seattle, WA 98103. Email: dgoldhaber@air.org • www.caldercenter.org

CALDER working papers have not undergone final formal review and should be cited as working papers. They are intended to encourage discussion and suggestions for revision before final publication. Any opinions, findings, and conclusions expressed in these papers are those of the authors and do not necessarily reflect the views of our funders. All errors are our own.

CALDER • American Institutes for Research
1400 Crystal Drive 10th Floor, Arlington, VA 22202
202-403-5796 • www.caldercenter.org

How Well Do Professional Reference Ratings Predict Teacher Performance?

Dan Goldhaber, Cyrus Grout, and Malcolm Wolff

CALDER Working Paper No. 272-1022
October 2022

Abstract

Most research about how to improve the teacher workforce has focused on interventions designed to improve incumbent teachers, far less attention has been directed toward teacher hiring processes and whether districts can make better hiring decisions. Using data from Spokane Public Schools and Washington state, we describe the findings from a study analyzing measures of the predictive validity of teacher applicant quality measures obtained from professional references. We find that professional reference ratings of prospective teachers are significantly predictive of teacher quality as measured by inservice performance evaluations and teacher value added in math. These findings are driven by applicants with at least some teaching experience and vary by rater type (e.g., principal or university supervisor); the magnitude of the relationship between the ratings of applicants and teacher performance is much smaller and not statistically significant for applicants that do not have teaching experience. Overall, the evidence suggests that obtaining explicit ratings of teacher applicants from professional references is a low-cost way to contribute to the applicant information available to hiring officials and has potential for improving hiring outcomes.

1. Introduction

Most research about how to improve the teacher workforce has focused on interventions designed to improve *inservice* teachers (through professional development or financial incentives, for example). Far less attention has focused on interventions targeted at *potential* teacher hires – teacher applicants. This gap in the literature is both surprising and problematic. The potential for improving workforce quality through effective hiring practices is broadly supported by research from the field of personnel economics (Heneman & Judge, 2003; Shaw & Lazear, 2007) and industrial psychology.[1] Making good hiring decisions is particularly important in the context of the teacher labor market. Once hired, it can be quite costly to remove a public school teacher who is ineffective (Gregory & Borland, 1999; National Council on Teacher Quality, 2014; *Vergara vs. State of California Tentative Decision*, 2014) and interventions designed to change the performance of incumbent teachers have been shown to have somewhat limited efficacy.[2]

School districts of course play a key role in influencing the composition of the teacher workforce.[3] In establishing their hiring processes, they specify the information applicants are required to provide, the design of screening and interview protocols, and how applicant information is used to inform hiring decisions.[4] One ubiquitous type of information collected by

[1] As Oyer and Schaefer (2011) note, "hiring the right employee is potentially as important or more so than motivating the employee to take the right action after the employee has been hired" (2011, p. 1772). See Society for Industrial and Organizational Psychology (2014) for an overview.

[2] Research generally suggests a weak to no relationship between these types of interventions to improve the performance of inservice teachers. See, for instance, Glazerman et al. (2010) and Wayne et al. (2008) on professional development and Glazerman and Seifullah (2010, 2012), Marsh et al. (2011), and Springer et al. (2010) on incentive pay (for more encouraging evidence on more comprehensive reforms that include both incentives, feedback, and professional development, see Dee and Wyckoff (2013) and Goldhaber and Walch (2012)). Low-performing novice teachers are unlikely to catch up with higher-performing peers (Atteberry et al., 2013)

[3] This is also true of states, which regulate the employment eligibility of prospective teachers through licensure systems (Goldhaber, 2011; Larsen et al., 2020).

[4] Contrary to conventional wisdom, districts often have a significant amount of choice among potential teachers. Two recent quantitative studies on teacher hiring, for instance, find the ratio of applicants to hires is over 7 to 1 (Goldhaber et al., 2017; Jacob et al., 2018). But while the overall ratio of teacher applicants to job openings is large

school districts is the letter of recommendation (Salgado, 2001). A limitation of utilizing letters of recommendation is that they can be time consuming to read and often lack candor, requiring hiring officials to read between the lines. In this paper, we examine the value of collecting more direct information about teacher applicants from those writing letters of recommendation. Specifically, we *analyze the degree to which categorical ratings of job applicants by their references are predictive of future job performance as teachers*. This research is of significant importance given the fact that teachers represent the largest public-sector occupation in the United States,[5] and that they have large impacts on both short-run and long-term student outcomes (e.g., Aaronson et al., 2007; Chetty et al., 2014c, 2014a; Jackson, 2018; Kraft, 2020; Rivkin et al., 2005).

We find a positive and significant relationship between reference ratings and both performance evaluations and teacher value-added in math. Receiving a rating in the top two categories relative to the bottom three categories predicts performance evaluation ratings that are 22% to 40% of a standard deviation higher. Similarly, receiving a rating in one of the top three categories relative to a rating in the bottom three categories is predictive of teacher value-added in math that is roughly 5% of a standard deviation higher (in student level standard deviations of test performance). The relationship between ratings and reading value-added is generally not statistically significant. These results are driven by applicants with some teaching experience; the magnitude of the relationship between the ratings of applicants and teacher performance is much smaller and not statistically significant for applicants that do not have prior teaching experience. Finally, we find evidence that the relationship between the letter writer and the applicant affects

in many districts, the applicant-job ratio tends to vary significantly across different subject areas and grade levels. For instance, there is evidence of persistent shortages of teachers in STEM and special education subject areas (Cowan et al., 2016; T. S. Dee & Goldhaber, 2017).

[5] See: https://www.bls.gov/oes/2020/may/public1.htm.

the predictive validity of the ratings: for instance, ratings of applicants from references identified as “Principal/Other Supervisor” or “Instructional Coach/Department Chair” tend to be more predictive of performance than ratings from other types of references, such as from a “Colleague”.

2. Background Literature and Professional Reference Ratings

2.1 Teacher Applicant Data and Teacher Performance

Until recently, the relationship between teacher characteristics typically available in administrative data and teacher performance was understood to be, at best, weak (Aaronson et al., 2007; Goldhaber, 2007a, 2007b; Rivkin et al., 2005). This led Staiger and Rockoff (2010), for instance, to the rather dreary conclusion that “School leaders have very little ability to select effective teachers during the initial hiring process” (p. 103). But the last decade has witnessed a number of new studies that provide more promising results about what can be learned about teachers through the hiring process. A chief reason for this is that a number of studies analyzed more refined information about new teachers and teacher applicants. Rockoff et al. (2011), for instance, collected measures of personality, cognitive ability, and content knowledge from first year math teachers in New York City. They found that these measures significantly add to the predictive validity that can be derived using the types of data about applicants that are usually available in administrative datasets (e.g., degree, major, passage of a teacher licensure exam).[6] While this supplemental information was collected from first-year math teachers rather than actual applicants, it’s collection could be integrated into the hiring process of a typical school district.

[6] The authors found that traditional and nontraditional information explained about 12% of the expected variance in teacher effectiveness compared to about 4% using only the types of data available from administrative datasets.

Several more recent studies have expanded the scope of research on teacher selection by using information collected during the hiring process. Jacob et al. (2018) studied applicants to Washington DC Public Schools (DCPS), who could opt to go through a centralized, multi-stage screening process to make it into a pool of "recommended" applicants. The process included evaluations of applicants based on their taking a written assessment of pedagogical and content knowledge, personal interviews, and teaching auditions.[7] The authors found that the applicant measures derived from the information collected during the screening process were significantly predictive of future performance as measured by a teacher's IMPACT score – a composite of observational performance evaluations, student progress measures, and (when available) teacher value-added.[8]

Bruno and Strunk (2019) studied whether applicant screening data collected by Los Angeles Unified School District (LAUSD) were predictive of outcomes for newly-hired teachers. The district's screening rubrics were used to score applicants on a structured interview, sample lesson, written responses to student-related scenarios, professional reference ratings, subject-area preparation, and academic background. The authors found that a composite screening score was predictive of subsequent performance as measured by teacher value-added in English language arts (ELA), observation-based performance evaluations, teacher attendance, and the propensity to stay in a school. They also found that individual elements of the screening process were differently predictive of teacher outcomes, pointing to potential tradeoffs under different screening strategies.

[7] Though not ultimately used for the purposes of screening under TeachDC, applicants also answered multiple choice questions from the Haberman Star Teacher Pre-Screener, a commercially available screening tool used by many urban school districts to assess applicant fit. Scores on the Pre-Screener were found to be significantly predictive of future performance.

[8] When the applicant measures are pooled into an index of predicted performance, the authors find a strong relationship with teacher IMPACT scores: the performance of teachers in the top quartile of predicted performance is 0.71 standard deviations higher than those in the bottom quartile.

Sajjadiani et al. (2019) used machine learning techniques to generate three measures of applicant quality using information provided by applicants to the Minneapolis Public School District: work experience relevance, tenure history, and attributions for previous turnover. Adopting a Heckman regression approach to account for potential sample selection bias, the authors found that these measures were predictive of teacher value-added, observation-based performance evaluations, student evaluations, and turnover.

Finally, and most closely related to our work described below, Goldhaber et al. (2017) analyzed the hiring process at Spokane Public Schools (SPS). They found that scores on a job-level applicant screening rubric used by SPS were significantly predictive of teacher value-added and retention. A one standard deviation increase in total screening score was associated with an increase in teacher value-added of between 0.03 and 0.07 standard deviations (measured at the student level) and a decrease in the propensity of attrition of 2.5 percentage points.

Letters of recommendation are an important source of information used by SPS hiring officials in rating applicants on the district's screening rubric. These letters, however, have to be interpreted by district hiring officials. The research we present in the current paper studies whether better information about job applicants can be solicited directly from their professional references. While the practice of obtaining information from job applicants' references is widespread (Salgado, 2001), the literature connecting this information to job performance is sparse.[9] This is an important omission. Incorporating the collection of references' ratings of applicants into the teacher application process is a potentially low-cost, easy-to-implement

[9] Mason et al. (2014) scored letters of recommendation (n = 41) from teacher candidates on a rubric and found that while scores were predictive being hired, they were not predictive of first-year performance as measured by principal evaluations. In contrast, Aamdot et al. (1993) found that positive traits identified letters of recommendation for applicants to a psychology graduate program were predictive of graduate GPAs and teaching ratings. Two studies that focus on the relationship between references' *ratings* of applicants and performance outcomes are limited by small sample sizes (50-60 applicants) and very specific job application settings – the Canadian military and graduate student internships at a single corporation (Liu et al., 2009; McCarthy & Goffin, 2001).

means of enhancing the applicant data available to hiring officials. This stands in contrast to the centralized screening systems studied by Jacob et al. (2018) and Bruno and Strunk (2019), which require one-on-one interactions with district administrators and can be quite costly.[10]

2.2 Collection and Properties of Professional Reference Ratings

The application process in SPS begins when an applicant creates a profile in the applicant management system used by the district. The profile includes information about the applicant's work experience, education, and credentials. Applicants are also asked to provide contact information for at least three professional references (PRs), who receive an e-mail from SPS directing them to submit a confidential letter of recommendation via an online submission form. The online form records each reference's contact information and relationship to the applicant.[11]

We began collecting structured assessments of applicants from their references in June 2015. Building on the system that prompts references to submit confidential letters of recommendation, references are redirected to an online survey following the submission of a letter. The survey form (see **Figure 1**) asks references to rate applicants relative to their peers on a series of criteria as follows: "Based on your professional experience, how do you rate this candidate relative to his/her peer group in terms of the following criteria?" The six criteria, which are described further in **Table B1** in Appendix B, are *Challenges Students*, *Classroom Management*, *Working with Diverse Groups of Students*, *Interpersonal Skills*, *Student Engagement*, and *Instructional Skills*.

[10] Jacob et al. (2018) estimated the total marginal cost of implementing the TeachDC system to be in the range of $70,000-$200,000 per year, or between $370-$1,170 per new hire.

[11] PRs indicate their relationship to applicants by selecting on of the following options: Principal, Assistant Principal, Principal Assistant, Supervisor, Director; University Supervisor; Instructional Coach, Department Chair; Supervising Teacher during student teacher placement; Colleague; Other.

The PR can rate the candidate as one of the following: *Among the best encountered in my Career (top 1%)*; *Outstanding (top 5%)*; *Excellent (top 10%)*; *Very good (well above average)*; *Average*; *Below Average*; *No basis for judgement*. Finally, the PR is asked to identify the competency in which the applicant is *Strongest*, the competency in which the applicant is *Weakest*, and to rate the applicant *Overall*.

As described in Goldhaber et al. (2021), "The relative percentile rating method is intended to solicit responses exhibiting enough variation across applicants for hiring officials to differentiate between strong and weak applicants (McCarthy & Goffin, 2001)" (p. 3). Applicants are likely to have a good relationship with their references, and we expect that references will want to describe applicants very positively or engage in "cheerleading" (as we discuss below, this is borne out in the ratings data we have collected). To accommodate this tendency, the ratings categories are concentrated at the top end of the distribution.

But despite efforts to discourage cheerleading in prior work (Goldhaber et al., 2021), we found that ratings were concentrated in the top three ratings categories (see **Figure A1** in Appendix A). On the *Overall* criterion, 22% of applicants were characterized as *Among the Best (top 1%)*, 35% as *Outstanding (top 5%)*, and 23% as *Excellent (top 10%)*. Hence, while the ratings appear to exaggerate applicant quality, they also exhibit a good deal of variation. The competencies raters identified as an applicant's strongest, or weakest, also exhibited a good deal of variation. The proportion of ratings identifying a competency as an applicant's strongest ranged from 7% (*Challenges Students*) to 23% (*Interpersonal Skills*), while the proportion of those identified as an applicant's weakest ranged from 6% (*Student Engagement*) to 26% (both *Challenges Students* and *Classroom Management*).

The degree of cheerleading was associated with the type of rater. Raters identified as an applicant's "Colleague" were the most generous in their ratings while those identified as the applicant's "Principal/Other Supervisor" were the harshest. For instance, principals rated 16% of applicants as *Among the Best (top 1%)* whereas colleagues rated 31% of applicants at that level.

The analysis of the inter-rater reliability of the reference ratings found that single-rater reliability ranged between 0.29 and 0.31 for summative ratings measures including the *Overall* criterion and the constructed measure *PR Factor*. We also found that inter-rater reliability was significantly higher for ratings of internal applicants versus external applicants, for experienced applicants versus novice applicants, and that ratings from raters identified as the applicant's *Principal/Other Supervisor* exhibited higher levels of reliability. Additionally, we found some evaluation criteria exhibited higher (*Classroom Management*) or lower (*Working with Diverse Groups of Students*) levels of reliability than other evaluation criteria.

The above information about the statistical properties of the PR ratings is important as it suggests that the *predictive validity* of the ratings (i.e., the extent to which we find they predict inservice teacher performance) may vary by applicant (internal/external) and rater types (principal, colleague, etc.) as well as the different ratings categories (classroom management, interpersonal skills, etc.).[12] Given this, our analyses (discussed in **Section 4**) analyze differences across these dimensions of applicants, raters, and ratings categories.

3. Data, Measures, and Analytic Sample

3.1 Data

To examine the link between PR ratings and subsequent outcomes for hired teachers we use five types of data: SPS applicant data, professional reference ratings data, SPS personnel

[12] For more discussion on this point, see Goldhaber et al., 2021.

data, statewide administrative data, and statewide student achievement data. We describe each of these in turn below.

Spokane applicant data come from the department of employment services at SPS. They maintain a record of each applicant and each application to a teaching position in the district in an electronic database using applicant tracking software. The data consists of the population of teacher applicants from 2015 to present, including applicants who were not ultimately hired (some of whom were SPS teachers who were applying for a new teaching position in the district). Applicant profiles contain information about applicants' education, experience, work history, personal statements, and letters of recommendation. The tracking of applicants is facilitated by an internal applicant ID number. Data on full names, certification numbers, and employee numbers allows applicant data to be linked to SPS personnel data and statewide administrative data. Job applications are identified by a job posting ID number and a status variable indicates whether the applicant was hired for the position. These records are linked to other applicant data using applicant names and to jobs by job posting numbers.

As noted above, we began collecting reference ratings data in mid-June 2015. Therefore, the ratings data provide partial coverage for letters of recommendation submitted by references in the 2015 hiring year and full coverage for applicants who created new applicant profiles for the 2016 to 2018 hiring years. An internal applicant ID used in the Spokane applicant data and the email address of the reference allow us to match each rating to the corresponding applicant profile and letter of recommendation. References redirected to the survey following the submission of a letter of recommendation submitted a ratings survey 95% of the time.

SPS personnel data are from the district's human resources department and include an employee ID number and information on teachers' names, positions, hire date, and ratings on the district's performance evaluation rubric. The performance evaluation data is based on a Marzano observational rating of teachers, known as the Teacher/Principal Evaluation Program (known in the state as "TPEP").[13] Teachers receive a comprehensive evaluation every three years that rates each teacher as belonging in one of four performance categories (Distinguished, Proficient, Basic, or Unsatisfactory) on eight different competencies: expectations, instruction, differentiation, content knowledge, learning environment, assessment, families and community, and professional practice. Teachers are evaluated annually if they are in a new job position or received rating of 1 or 2 on one or more evaluation criteria.[14]

Evaluation ratings on each competency are available for job applicants observed teaching in Spokane but are unobserved for applicants who are not employed in SPS. TPEP ratings are largely based on classroom observations, which are regularly used to judge teacher quality and have been shown to be correlated with student achievement gains (e.g., Kane, Taylor, Tyler, & Wooten, 2011). Similar to Jacob et al. (2018) who report a correlation of 0.27, we find correlations between teachers' summative TPEP ratings and teacher value-added of 0.23 (math) and 0.21 (reading).

We obtain statewide administrative data from two sources: the Washington State S-275 personnel reporting system and the Comprehensive Education Data and Research System (CEDARS**)** database. The S-275 data are compiled by the Office of the Superintendent of Public Instruction (OSPI) and report information about all school district personnel under contract as of October 1 of each school year, including position assignments, compensation, experience,

[13] SPS rolled out the adoption of a Marzano-based teacher evaluation rubric during the 2013-14 to 2015-16 school years as part of a statewide initiative to improve the utility of districts' teacher evaluation protocols which, had historically generated ratings with very little variation.
[14]For more detail about Washington State's TPEP evaluation program see https://www.k12.wa.us/educator-support/teacherprincipal-evaluation-program (accessed June 6, 2022).

highest degree held, ethnicity, and gender. Records of full names and unique certificate IDs facilitate the matching of these data to the district-level data from SPS. OPSI also provides teacher licensure data, including scores on licensure exams and subject-area endorsements. The CEDARS database includes unique classroom IDs for both teachers and students, allowing them to be linked together.[15] It also includes information on individual students' backgrounds, including gender, race/ethnicity, free or reduced-priced lunch, migrant, and homeless statuses, as well as participation in the following programs: home based learning, learning disabled, gifted/highly capable, limited English proficiency (LEP), and special education.

3.2 Measures

Here we describe the measures we will use to model the relationship between reference ratings and teacher performance.

Reference Ratings

The ratings solicited from applicants' references are ordered categorical data, and we can model the relationship between each rating criterion and teacher performance in a regression context by representing the ratings as categorical variables. To facilitate the analysis, we also fit a graded response model to the six ratings criteria to generate a summative measure of the ratings provided by each reference rating submission. The graded response model is suited to ordered categorical data and allows items (in this case, the different ratings criteria) to vary by difficulty and discrimination.[16] Following Chen et al. (2021), we specify the following model where the probability of observing rating level k or higher for criterion j and applicant rating i is given by:

[15] CEDARS data includes fields designed to link students to their individual teachers, based on reported schedules. However, limitations of reporting standards and practices across the state may result in ambiguities or inaccuracies around these links.

[16] A naïve approach to generating a summative rating measure might assign numerical values to each ratings category and calculate an average rating across criteria. This would assume, however, that (1) receiving a particular rating on one criterion was equivalent to receiving that rating on each of the other criteria (i.e., that the *difficulty* of the ratings criteria does not vary), and (2) that the difference between receiving a low vs. high rating on a particular

(1) $$\Pr(Y_{ij} \geq k | \theta_i) = \frac{\exp\{a_j(\theta_i - b_{jk})\}}{1+\exp\{a_j(\theta_i - b_{jk})\}},$$

where a_j represents the discrimination of criterion j, b_{jk} is the kth cutpoint of criterion j, and θ_i is the latent quality of applicant i.[17]

Performance Evaluation Ratings

As noted above in **Section 3.1**, teachers receive performance evaluations on which they are rated as belonging in one of four performance categories on eight different competencies. To generate a summative measure of teacher performance, we apply the graded response model described in equation (1) to these evaluation criteria, using the empirical Bayes estimate of θ_i as a summative measure of each teacher's performance evaluation.

To account for the fact that teachers are evaluated by different raters, who may assess teachers' performance more or less harshly, we estimate the following linear regression model:

(2) $$\hat{\theta}_{it} = \gamma_{st} + \varepsilon_{it},$$

where $\hat{\theta}_{it}$ is the summative measure of teacher i's performance evaluation in year t and γ_{st} is a school-year fixed effect.[18] Following Cowan et al. (2022), we use the residuals from this model as our teacher performance evaluation measure in the predictive validity models described below in **Section 4.1**.

Teacher Value-Added

criterion is equivalent to receiving a low vs. high rating on each of the other criteria (i.e., that the *discrimination* of the ratings criteria does not vary).

[17] We estimate this model using Stata's *irt grm* program and use the empirical Bayes estimates of θ_i as a summative measure of each applicant rating. We cannot calculate the empirical Bayes estimates of θ_i for reference ratings for which one or more criteria are rated as "No basis for judgement".

[18] Unfortunately, we do not have consistent data regarding who is each teacher's evaluator, and for this reason adopt a school-year fixed effect rather than an evaluator-year fixed effect. We also estimate specifications that include controls for classroom characteristics (but which are available for only a subset of teachers), which have been shown to influence teachers' performance evaluation ratings (Steinberg & Garrett, 2016) and find very similar results.

We model student achievement to obtain estimates of teacher value-added in math and reading separately. Consistent with Goldhaber et al. (Goldhaber et al., 2018), for teachers who teach in tested grades and subjects, we estimate each teacher's value-added effectiveness based on the following model:

$$Y_{ijst} = \alpha_{jst} + Y_{is(t-1)}^{\top}\delta + X_{it}^{\top}\gamma + \varepsilon_{ijt}, \quad (4)$$

Where Y_{ijst} is the state test score for each student i with teacher j in subject s (math or reading) and year t, normalized within grade and year. $Y_{is(t-1)}$ is a vector of prior test scores in subjects s (both math and reading), X_{it} is a vector of student characteristics in year t (gender, race, economically disadvantaged status, English language learner status, gifted status, special education status, learning disability status), and α_{jst} is a teacher fixed effect in subject s and year t.[19] We use the estimates of α_{jst} as our measure of teacher value-added to student achievement.

3.3 Analytic Sample

Our analytic sample is anchored by the population of individuals who applied for one or more teaching positions in SPS during the 2015 to 2018 hiring years and for whom we collected one or more reference ratings. We observe inservice outcomes for a subset of applicants who are employed by SPS (performance evaluation ratings) or by public schools in Washington (teacher value-added) during the 2015-16 to 2018-2019 school years.

As shown in **Table 1**, we observe ratings of 3,588 applicants during these hiring years, corresponding to a total of 11,678 reference ratings of applicants who applied to one or more teaching positions in SPS.[20] We observe different types of teacher performance outcomes for

[19] Student characteristics include average prior achievement in math and reading, proportion female, American Indian, Asian or Pacific Islander, Black, Hispanic, and proportion of students with learning disabilities, gifted status, and free and reduced-price lunch status.

[20] Note that we treat individuals who apply for a position in different hiring years as different "unique applicants."

teachers employed in a Washington State public school subsequent to their application to SPS. For teachers hired in SPS we observe inservice performance evaluations (see **Section 3.1**) and for the subset in tested grades and subjects (grades 4-8), we have value-added measures of performance. For those hired into districts outside of SPS we can also estimate value-added for the subset in tested grades and subjects, but we do not have district observational ratings.

We observe 914 applicants subsequently employed in a classroom teaching position in SPS.[21] Among these applicants, 576 were newly employed by SPS in a classroom teaching position (column (2)), 119 were employed as a classroom teacher in the prior year in a different building, and 219 teachers remained in the same building as in the prior year. We also observe 813 applicants who were subsequently employed in classroom teaching positions by other Washington State school districts.

As discussed in **Section 2.2**, reference ratings skew toward to the top end of the distribution. In our study sample 54% of ratings indicate an applicant is either *Among the Best (top 1%)* or *Outstanding (top 5%)*, with the ratings of subsequently employed applicants tracking slightly higher. Among applicants employed by SPS, those percentages are higher – between 58% (column 3) and 60% (column 2). The summative rating measure *GRM* derived in **Section 4.2** is 5% to 11% of a standard deviation higher among applicants subsequently employed in SPS (columns (2) to (4)) relative to all applicants and 12% to 18% of a standard deviation higher relative to applicants who are not subsequently employed (column (6)).

The most common rater type is *Principal/Other Supervisor* (35%) followed by *Colleague* (26%). Subsequently employed applicants are slightly more likely to have a rating from a

[21] An applicant's status as being employed in a classroom teaching position is based on their appearing in the S-275 administrative records maintained by the state with a duty code classification of 31, 32, 33, 34. This excludes applicants employed by the district in other certificated positions such as librarian or subject area coach.

Principal/Other Supervisor and slightly less likely to have a rating from a rater identified as *Colleague* or *Other.* Among applicants we are able to link to S-275 administrative records, 30% had no prior teaching experience and nearly a quarter had 6 or more years of experience.

Overall, we are able to link 23% of reference ratings to subsequent performance evaluations, including 91% of ratings where the applicant was newly observed in a classroom teaching position in SPS, 61% of ratings where the applicant transferred to a new teaching position within SPS, and 50% of ratings where the applicant remained the same position.[22] We observe teacher value-added outcomes for smaller percentages of applicants – 7% overall, and between 8% and 14% of ratings of applicants subsequently employed by SPS and between 15% and 18% of ratings of applicants subsequently employed by other school districts.

4. Empirical Methods

Our analyses address the question, to what extent are reference ratings predictive of subsequent performance? We outline our approach to answering this question and addressing potential bias, introduced by selection into the sample, below.

4.1 Predictive Validity

To examine the relationship between professional reference ratings and subsequent performance we consider two measures of performance: teacher ratings on the evaluation instrument – the Teacher/Principal Evaluation Program (TPEP) – and teacher value-added.

<u>Teacher Performance Evaluations</u>

[22] We are able to link 96% of ratings of teachers newly observed in a classroom teaching position to performance evaluation records. Experienced employees in good standing receive comprehensive evaluations every three years. An employee newly observed in a classroom teaching position might not have a comprehensive evaluation rating if they were in previously in a different type of certificated position and in good standing or if they left their position or went on leave mid-year.

Using the summative performance evaluation measure obtained above, $TPEP_{jt}$, we estimate the following linear regression model with standard errors clustered at the applicant-year level:

$$(5) \qquad TPEP_{jt} = \beta_0 + RR_{jrt}\beta_1 + Y_{jt}^{\top}\beta_2 + \varepsilon_{jt}.$$

The variable of interest is RR_{jrt} (which is either the reference rating represented as a categorical variable, or the summative ratings measure derived above) for teacher *j* and rater *r* prior to evaluation year *t*. In some specifications we also control for a vector of teacher characteristics Y_{jt} that could potentially influence TPEP ratings including ethnicity, gender, whether a teacher holds an advanced degree, and teacher experience.[23] To examine the possibility that reference ratings are differentially predictive for novice and experienced applicants, we also estimate model (5) separately for these two groups.[24]

Finally, different types of professional references may be differentially able to assess applicants' teaching ability, or be more or less likely to engage in "cheerleading" (Leising et al., 2010). Recall, for instance, that professional references identified as Principal/Other Supervisor are less likely than other professional references to rate applicants as "Top 1%" and that their ratings exhibited higher levels of inter-rater reliability (Goldhaber et al. 2021). To test for differential predictive validity across rater type, we modify equation (5) above as follows:

$$(6) \qquad TPEP_{jt} = \beta_0 + \left(GRM_{jrt} \times R_{jrt}\right)^{\top}\beta_1 + R_{jrt}^{\top}\beta_2 + T_{jrt}\beta_3 + Y_{jt}^{\top}\beta_4 + \varepsilon_{jt},$$

[23] Employee characteristics have been found to influence judgements of performance in other professions, including gender and ethnicity (Greenhaus et al., 1990; Grissom & Bartanen, 2022). Regarding experience, the Marzano performance evaluation framework used by SPS is growth oriented and beginning teachers are expected to score lower than experienced teachers. Additionally, returns to experience – as measured by both value-added and performance evaluations – are well documented and it is possible that TPEP evaluators' knowledge of this may also influence their assessments of teacher performance.

[24] An applicant is considered to be a novice if they do not report prior teaching experience in their job application profile.

where R_{jrt} is the relationship between rater r and teacher j, and the interaction term $(GRM_{jrt} \times R_{jrt})$ allows the slope coefficient β_1 to vary according to rater type.

Teacher Value-Added

We assess the predictive validity of the reference ratings for teacher value-added using the teacher fixed-effects, $\hat{\alpha}_{jst}$, estimated in model (4) above. We estimate the following model:

$$(7) \qquad \hat{\alpha}_{jst} = \beta_0 + GRM_{jrt}\beta_1 + Y_{jt}^{\mathsf{T}}\beta_2 + T_{jrt} + \varepsilon_{jt},$$

where $\hat{\alpha}_{jst}$ is the single year value added for teacher j in subject s (i.e., math or reading).[25] Standard errors are clustered at the applicant-year level. As with annual performance evaluations, to account for the possibility that reference ratings are differentially predictive for novice and experienced applicants, we estimate model (7) separately for these two groups. We also estimate a specification with the rater type fixed effects and interaction term $GRM_{jrt} \times R_{jrt}$ that allows the slope coefficient β_1 to vary according to rater type.

4.2 Accounting for Sample Selection

The models described in **Section 4.1** estimate the extent to which reference ratings are predictive of teacher performance. Importantly, as noted above, we only observe TPEP ratings when an applicant is hired by SPS or remains employed in a previously held teaching position in SPS, and we only observe teacher value-added when an applicant is hired by SPS or a different public school district in Washington State *and* teaches in a tested grade and subject area. This raises the concern that our findings may suffer from selection bias. In particular, we might expect a naïve assessment of the relationship between reference ratings and teacher outcomes to be

[25] We also estimate specifications that include categorical variable for teacher experience Y_{jt}. This is to account for the fact that professional references are asked to rate teacher applicants relative to their peers and they may view the peer group to be teachers with similar experience. The findings are not sensitive to the inclusion of teacher experience.

biased downward as applicants hired despite poor recommendations are likely to have unobserved attributes that make them desirable hires.

We address the potential for selection bias by estimating a Heckman selection model (Heckman, 1979) to account for the likelihood of observing the teachers in the sample of hired applicants. To employ this technique, we utilize two variables related to competition for jobs that are arguably, and by our assumption, predictive of hiring outcomes and uncorrelated with applicants' subsequent outcomes:

1) *Quantity of competition* – The number of unique applicants against which each applicant is competing divided by the number of job openings they have applied for.[26]
2) *Quality of competition* – The average quality of the applicants against which each applicant is competing (those who applied to one or more of the same job postings) as measured by the average of the mean summative reference rating of the competing applicants.

Cursory evidence suggests the level of competition for a teaching position in SPS varies a great deal according to school level and subject area. For instance, in our study sample, non-special education elementary job postings averaged 49 applications whereas the average secondary math job posting drew 10 applications.

To implement the first stage of the Heckman model, we estimate the following probit model of an applicant's propensity to be hired and have an observed measure of performance using the job competition variables described above, Z_{jt}, as instruments:

[26] For example, suppose that an applicant applied for six job postings that corresponded to nine openings, and 100 other applicants also applied to one of more of those job postings. That applicant's *quantity of competition* would be $100/9$.

(8) $$Hired^{*} = \beta_0 + GRM_{jrt}\beta_1 + A_{jt}^{\top}\beta_2 + R_{jrt}^{\top}\beta_3 + Z_{jt}^{\top}\beta_4 + \varepsilon_{jt},$$

where $Hired^{*}$ is equal to 1 if an applicant is hired for one of the positions to which they applied and has an observed measure of performance for the outcome in question (performance evaluation, value-added in math, or value-added in reading).[27] The first-stage sample is restricted to those applicants who applied to one or more job postings for which an outcome was observed for the hired applicant.[28] We include a vector of applicant characteristics available to SPS during the hiring process, A_{jt}, including a categorical control for experience, an indicator for whether the applicant reports holding an advanced degree, indicators for applicant ethnicity, and a categorical control for applicant-rater relationship type. Then, letting

(9) $$Hired = I(Hired^{*} \geq 0),$$

we estimate the conditional model

(10) $$TPEP_{jt} \; or \; \hat{\alpha}_{jst} |\, Hired = \beta_0 + GRM_{jrt}\beta_1 + A_{jt}^{\top}\beta_2 + R_{jrt}^{\top}\beta_3 + \lambda_{\mathrm{jt}}\beta_4 + \varepsilon_{jt},$$

where λ_{jt} is the Inverse Mill's Ratio calculated based on the first-stage estimates.[29]

5. Results

In **Section 5.1** we examine the degree to which reference ratings are predictive of teacher ("TPEP") evaluation performance and in 5.2 we assess the predictive validity of reference ratings for teacher value-added. We address the potential for bias stemming from sample selection in **Section 5.3**.

[27] Existing SPS teachers who apply for a new job within the district and are hired are coded as hired and those who are not hired into a *new* position but remain employed by SPS are coded as not hired. But, as we note below, the primary findings are similar to estimates that rely solely on applicants that are new (i.e., not incumbent SPS teachers) to SPS.

[28] For example, an applicant who only applied to secondary math positions would not be included in the first stage of the reading model because no reading outcomes would be observed for the applicants hired into those positions.

[29] The model is estimated as a two-step model with bootstrapped standard errors clustered at the applicant-year level using Stata's *Heckman* command.

5.1 PR Ratings and ("TPEP") Performance Evaluation Ratings

Results on the relationship between the summative ratings measure *GRM* and the performance evaluation measure derived above in **Section 3.2** are presented in columns (1) and (2) of **Table 2**. Column (1) includes *GRM* alone and column (2) adds a vector of controls for teacher characteristics described in **Section 4.1**. We find that *GRM* is significantly predictive of performance evaluation ratings, with a one standard deviation change in *GRM* associated with a 13% to 14% standard deviation change in performance.[30] Columns (3) and (4) of **Table 2** present the same pair of regressions with the rating criterion *Overall* represented as a categorical variable. Here, we find effect sizes that are quite large, with applicants scoring in the top two ratings categories predicted to perform between 21% and 40% of a standard deviation higher than applicants rated *Very Good*. These findings are robust to the inclusion of controls for teacher characteristics, including indicators for whether they are female, white, or hold an advanced degree, and a categorical control for experience.[31]

As noted above, in prior work (Goldhaber et al., 2021) we found that the inter-rater reliability of the reference ratings was lower among novice applicants than among experienced applicants, and inter-rater reliability is a limiting factor for the predictive validity of the reference ratings. In **Table 3**, we replicate models presented in **Table 2**, but split the sample into ratings of novice applicants (Panel A) and experienced applicants (Panel B).[32] We find that reference ratings of novice applicants are *not* significantly predictive of subsequent performance

[30] This is comparable in magnitude to findings reported by Sajjadiani et al. (2019) regarding their measure of tenure history but smaller than both applicant measures used in Jacob et al. (2018) and Bruno and Strunk (2019).
[31] While more detailed information on teacher race and ethnicity is available, we adopt the binary white/non-white indicator to maintain cell sizes greater than 10 observations.
[32] An applicants' novice status is determined by whether the applicant reports prior teaching experience in their application profile.

evaluations. Conversely, ratings of experienced applicants are more strongly predictive of performance evaluations than for the ratings of applicants in general (as shown in **Table 2**).[33]

In **Table 4**, we consider whether ratings from different types of raters (as defined by the rater's relationship to the applicant) are more or less predictive of performance evaluations. We begin (in columns 1 and 2) by examining whether *receiving* a rating from a particular type of rater is predictive of performance (regardless of the rating that is received); this is of interest in that school systems sometimes require ratings from particular types of references. In general, there are no statistically significant differences between the various rater types (the reference category is colleagues), though the estimates are not terribly precise. The exception in the sparse specification (column 1) is receiving a rating from a cooperating teacher, which predicts lower performance, but this relationship is no longer significant when controlling for teacher characteristics. It is also important to note that the type of individual that applicants call upon for a reference is likely to be related to the applicant's career stage and perhaps other unmeasured applicant attributes.[34]

In columns (3) and (4) we include the rater-type fixed effects to account for the fact that different rater types tend to rate more or less generously (see **Figure A1**) and allow the coefficient on the summative rating measure *GRM* to vary according to rater type. As indicated by the interacted slope coefficients, we find that ratings of applicants from references identified as *Principal/Other Supervisor*, *Instructional Coach or Department Chair*, or *Colleague* are strongly predictive of subsequent performance evaluations. However, ratings from other types of

[33] This is verified by estimating a pooled model that includes an interaction between an indicator for *experienced* and the *GRM* variable.

[34] Among novice applicants in the regression sample, 54% of reference ratings are from references identified as the applicants' *Cooperating Teacher* or *University Supervisor*. Among experienced applicants this is true of only 14% of reference ratings.

raters – those identified as an applicant's *Cooperating Teacher*, *University Supervisor*, or as *Other*, are not significantly predictive of performance. This finding is further illustrated in **Figure 2**, which presents predicted levels of performance according to summative rating level and rater type. We can see that low vs. high ratings from references identified as *Principal/Other Supervisor*, *Instructional Coach or Department Chair*, or *Colleague* predict substantially different levels of performance, whereas ratings from other types of raters predict about the same level of performance regardless of the rating level.

The series of results presented above analyze two different summative measures of applicant ratings – the summative *GRM* measure and ratings on the *Overall* criterion. In **Table 5** we consider each evaluation criterion in turn. We find that for each of the evaluation criteria, a rating of *Among the Best (top 1%)* is significantly predictive of a higher performance evaluation rating relative to the reference category of *Very Good*. And for four of the evaluation criteria, the same is true of receiving a rating of *Outstanding (top 5%)* – the exceptions being the criteria *Classroom Management* and *Interpersonal Skills*. While each of the evaluation criteria is significantly predictive of performance, we find substantial variation in magnitude. Relative to receiving a rating of *Very Good*, receiving a top rating on the *Interpersonal Skills* criterion is predictive of a performance evaluation rating that is 21% of a standard deviation higher, whereas a top rating on the *Instructional Skills* criterion is predictive of performance that is 45% of a standard deviation higher. Receiving a top rating on the other evaluation criteria is similarly predictive of higher performance – between 30% and 35% of a standard deviation. Another way to assess the degree to which ratings on the different criteria explain is to compare the R-squared statistics of the regressions. By that measure, the *Classroom Management* and *Instructional*

Skills criteria perform the best while the *Challenges Students* and *Working with Diverse Groups of Students* criteria exhibit the least explanatory power.[35]

In addition to rating applicants on a series of competencies relative to their peers, applicants' references are also asked to select the competencies in which the applicant is *Strongest* and *Weakest* (see **Figure 1**). As discussed in Goldhaber et al. (2021), two competencies that stand out are *Challenges Students* and *Classroom Management*, which are identified as an applicant's strongest competency much less frequently than are other competencies, and much more frequently as an applicant's weakest competency. However, as shown in **Table 6**, when we regress our performance evaluation measure on a vector of indicators for the competency identified as an applicant's *Strongest*, we fail to find any relationship to performance. The same is true when we estimate a separate regression model with indicators for the competency identified as an applicant's *Weakest.*

5.2 PR Ratings and Teacher Value-Added

Here, we examine the relationship between reference ratings and the teacher value-added measures derived in **Section 3.2**,[36] presenting the same series of analyses as for the performance evaluation outcomes above.

Results on the relationship between the summative ratings measure *GRM* and value-added in math is presented in columns (1) and (2) of **Table 7**, with the rating on the *Overall* criterion represented as a categorical variable presented in columns (3) and (4). The equivalent set of results for value-added in reading is presented in columns (5) to (8). We find that *GRM* is significantly predictive of value-added in math, with a one-standard deviation change in *GRM*

[35] F-Tests show that as a each of the individual criterion regression is statistically significant.
[36] Note that the value-added measures are standardized in student achievement terms.

associated with a 2% of a standard deviation change in teacher value-added. Similarly, receiving a rating on the *Overall* criterion in the top three ratings categories is predictive of an increase in value-added in math between 4 and 7% of a standard deviation relative to receiving a rating of *Very Good*, but there is little differentiation within the top three ratings categories. These results are robust to the inclusion of a categorical control for teacher experience. The relationship between the ratings and value-added in reading is positive but not statistically significant, and due to the lack of precision, we can't rule out either a slightly negative or modestly positive relationship.

As with performance evaluations, we examine how predictive validity varies according to whether an applicant is a novice or experienced, splitting the sample into ratings of novice applicants (Panel A) and experienced applicants (Panel B) in **Table 8**. Consistent with the findings of performance evaluation outcomes, we find that the summative rating measure *GRM* is predictive of value-added in math for experienced applicants but not for novice applicants. But small samples of novices and the corresponding lack of precision preclude us from drawing strong conclusions.

When we estimate a pooled model with a novice/experienced interaction term on the *GRM* variable, the difference between the estimated coefficients on *GRM* are not statistically significant. And while the magnitude of the coefficients on the ratings categories of the *Overall* criterion are slightly larger for experienced applicants, they are not statistically significant for either group. We also do not find evidence that reference ratings are significantly predictive of value-added in reading for either group though, here too, our inferences are limited by small sample sizes and a corresponding lack of precision.

In **Table 9,** we consider whether the relationship between reference ratings and teacher value-added varies according to rater type. The models in this table include either a rater type fixed effect, showing how predictive it is to have a reference from a particular rater type, or both a rater type fixed effect and an interaction between the rater type and the summative rating measure (*GRM*), showing whether the predictive power of changes in ratings vary by rater type. We find that receiving a rating from a colleague is positively predictive of value-added in both math and reading relative to receiving a rating from a different type of rating, a result that contrasts with our findings on performance evaluations.[37] Regarding the interacted slope coefficients on *GRM*, we find that ratings from raters identified as an applicant's *Principal/Other Supervisor*, and *Instructional Coach or Department Chair* are significantly predictive of value-added in math, and ratings from those identified as *Instructional Coach/Department Chair* are predictive of value-added in reading.[38] These findings are consistent with the findings for performance evaluation outcomes in that ratings from references identified as an applicant's *Cooperating Teacher*, *University Supervisor*, or as *Other* are not predictive of performance. However, while ratings from references identified as *Colleagues* are predictive of performance evaluation outcomes, they are not predictive of teacher value-added.

As above in regard to performance evaluations, we further illustrate how predictive validity varies according to rater type by presenting predictive levels of value-added according to summative rater level and rater type. In **Figure 3**, we can see that receiving a low vs. high summative rating from a reference identified as a *Principal/Other Supervisor* or *Instructional*

[37] Though again it is important to recall that the type of rater from whom applicants get recommendations may be related to unmeasured applicant characteristics.

[38] While the findings on raters identified as *Instruction Coach/Department Chair* are promising, it is worth noting that they account for only a small number of ratings – 36 observations in the Math models and 44 observations in the reading sample.

Coach/Department Chair predicts substantially different levels of value-added math, whereas ratings from other types of raters do not differentially predict levels of performance. In **Figure 4**, we see that only ratings from references identified as *Instructional Coach/Department Chair* predict significantly different levels of value-added in reading, and that these predictions are fairly imprecise.

Considering each evaluation criterion in turn in **Table 10**, we find that for four of the six evaluation criteria, a rating of *Among the Best (top 1%)* is significantly predictive of a higher teacher value-added in math relative to the reference category of *Very Good*, the exceptions being the criteria *Working with Diverse Groups of Students* and *Interpersonal Skills*. Consistent with our findings on performance evaluation outcomes, we find that the *Instructional Skills* criterion is most strongly predictive of value-added in math and that the *Interpersonal Skills* criterion has the weakest relationship. The results are less consistent for value-added in reading, for which receiving a top rating is not generally predictive of performance, an exception being the *Classroom Management* criterion. As noted above, one might also assess how well the ratings on the different criteria predict performance by comparing R-squared statistics from the different regressions. By that measure, the *Classroom Management*, *Student Engagement*, and *Instructional Skills* criteria perform the best for both math and reading and the *Interpersonal Skills* criterion performs the least well.[39] As noted above in **Section 2.2** prior work showed that the *Classroom Management* criterion exhibited a higher – and the *Working with Diverse Groups of Students* criterion a lower – level of inter-rater reliability than other criteria.

[39] F-Tests show that the regressions for the criteria *Classroom Management, Student Engagement,* and *Instructional Skills* statistically significant for value-added in math, and that only the regression for *Classroom Management* is statistically significant for reading.

Turning to the criteria in which applicants are rated as being *Strongest* or *Weakest* in **Table 11**, we find little relationship between these ratings and performance as measured by either value-added in math or reading. While the criterion *Classroom Management* being identified as an applicant's *Strongest* competency is significantly predictive of value-added in math relative to the reference category (*Interpersonal Skills*), it is not significantly different from the other categories. This lack of any relationship between the *Strongest*/*Weakest* ratings and value-added is consistent with our findings for performance evaluation ratings.

5.3 Addressing Sample Selection

We assess whether the above findings may suffer from sample selection bias by estimating the Heckman selection models described in **Section 4.2**. The results from the selection models are presented in **Table 12** (performance evaluation outcomes) and **Table 13** (value-added in math and reading).

Regarding performance evaluations, we find that the summative reference ratings measure *GRM* is predictive of selection into the sample (i.e., being hired for a position that was applied for and having an observed performance evaluation in the subsequent school year). The two instruments – *quantity of competition* and *quality of competition* – are both of the expected sign, but only the *quantity of competition* measure is statistically significant. The coefficient on the Inverse Mills Ratio (*Lambda* in column (2)) in the second stage of the model is statistically insignificant, suggesting that the findings are not significantly biased by selection into the sample. For the purpose of comparison, in column (3) we present results for an uncorrected model (*Lambda* is excluded) run on the same set of observations and find that the magnitude of the coefficient on *GRM* is somewhat smaller in the uncorrected model (0.168 vs. 0.222).

Regarding teacher value-added, we find that the summative reference ratings measure *GRM* is not predictive of selection into the sample for value-added in math, and that it *is* predictive for value-added in reading. As shown in **Table 13**, in both cases the instrument *quantity of competition* is of the expected sign and statistically significant whereas the *quality of competition* instrument is statistically significant and positive. This latter result is puzzling in that it suggests that applicants who face competition of higher average quality are *more* likely to be hired. For both math and reading, the coefficient on the Inverse Mills Ratio is statistically insignificant, again suggesting that our findings are not biased by selection into the sample. That said, given the small sample size and the behavior of the instrumental variables, we are unable to draw any strong conclusion from the evidence on sample selection for the teacher value-added.

6. Discussion, Policy Implications and Conclusions

In a recent analysis of the teacher labor market in Boston, James et al. (2022) found that the number of applicants to a teaching position is largely unassociated with the quality of the hire, suggesting that 1) schools struggle to identify the best candidates, and 2) there is an unrealized potential for improving teacher hiring. As the authors note, "Districts can take steps to improve teacher quality through the hiring process, but without improved screening and selection these efforts will fail to realize their full potential." (p. 27).

The good news is that a growing body of evidence demonstrates that teacher applicant information collected during the hiring process is predictive of subsequent teacher outcomes (Bruno & Strunk, 2018; Goldhaber et al., 2017; Jacob et al., 2018; Sajjadiani et al., 2018). However, the process for collecting and evaluating this information can be costly. For instance, Jacob et al. (2018) estimated the total marginal cost of implementing the TeachDC system to be in the range of $70,000-$200,000 per year, or between $370-$1,170 per new hire.

An alternative, low-cost way of deriving information about teacher applicants that we explore here is asking professional references to rate applicants on criteria expected to be predictive of subsequent performance. Yet, there is surprisingly little evidence about the degree to which references are able to predict an employee's future performance, despite the fact that the practice of asking job applicants to provide recommendations from professional references is nearly ubiquitous.

The potential for the collection of reference ratings to improve hiring outcomes depends on a number of factors, including how hiring officials utilize the information provided by reference ratings and the extent to which the ratings provide information that is supplemental to (as opposed to overlapping with) existing applicant information. But perhaps most fundamentally, it depends on the degree to which reference ratings are predictive of subsequent performance, the focus of this paper. On this front, our findings are promising but also demonstrate some limitations in the potential for reference ratings to inform hiring decisions.

We find an overall positive and significant relationship between reference ratings and subsequent performance evaluations of teachers employed by SPS and evidence of a modest relationship between reference ratings and teacher value-added in math. The strength of these relationships that we find between reference ratings and teacher performance is generally smaller than that found in previous studies, which analyzed more intensive applicant screening mechanisms. For instance, in prior work studying an applicant screening rubric scored by school-level hiring officials (Goldhaber et. al, 2017), we found effect sizes that were roughly three times as large for teacher value-added in math. Similarly, in studying applicant screening score index utilized by Washington, DC Public Schools, Jacob et al. (2018) found a relationship to performance evaluation ratings that was roughly 1.5 times as large. But it is important to note that the ratings we analyze are based on a very inexpensive survey of professional references as opposed to the more intensive means of assessing teacher applicants described in the aforementioned studies.

We also find that the strength of the relationship between reference ratings and teacher performance varies according to both rater and applicant characteristics. Specifically, the signal that the reference ratings send on applicant quality appears to be driven by certain types of raters: references identified as an applicant's *Principal/Other Supervisor*, *Instructional Coach/Dept. Chair, or Colleague.* The finding on principals is consistent with earlier research showing that principals can generally identify which teachers are most likely to produce the largest gains in student achievement (Jacob & Lefgren, 2008). And, relatedly, the strength of the relationship between applicant ratings and future teacher performance is far stronger among applicants with prior teaching experience; we fail to find significant evidence of a relationship between ratings and subsequent performance among novice applicants.

That the relationship is strong for incumbent teachers is not surprising as there is strong evidence that being able to observe teachers in practice is predictive of their future performance (e.g., Kane et al., 2011). Novice applicants may need to be evaluated on a different set of criteria, or it may simply be difficult to assess the ability of a novice applicant given their lack of experience leading a classroom. Indeed, in prior work we found that ratings of novice applicants demonstrated lower levels of inter-rater reliability, which dampens the predictive validity of a measure (Goldhaber et al., 2021).

Overall, our findings show that meaningful information can be solicited from professional references in the form of categorical ratings of applicants. But they point to the need to learn more about teacher applicants who have not previously participating in the labor market as public school teachers.

References

Aamodt, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting Performance with Letters of Recommendation. *Public Personnel Management*, *22*(1), 81–90. https://doi.org/10.1177/009102609302200106

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135. https://doi.org/10.1086/508733

Atteberry, A., Loeb, S., & Wyckoff, J. (2013). *Do First Impressions Matter? Improvement in Early Career Teacher Effectiveness* (Issue CALDER Working Paper 90). http://www.nber.org/papers/w19096

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The Cost of Teacher Turnover in Five School Districts: A Pilot Study*. National Commission on Teaching and America's Future.

Bruno, P., & Strunk, K. O. (2019). Making the Cut: The Effectiveness of Teacher Screening and Hiring in the Los Angeles Unified School District. *Educational Evaluation and Policy Analysis*, *41*(4), 426–460. https://doi.org/10.3102/0162373719865561

Chen, B., Cowan, J., Goldhaber, D., & Theobald, R. (2021). *From the Clinical Experience to the Classroom : Assessing the Predictive Validity of the Massachusetts Candidate Assessment of Performance* (No. 223-1019–2; CALDER Working Paper). https://caldercenter.org/sites/default/files/WP 223-1019-2.pdf

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). The Long-Term Impacts Of Teachers: Teacher Value-Added And Student Outcomes In Adulthood. *American Economic Review*, *104*(9), 2633–2679. http://www.nber.org/papers/w17699

Chetty, R., Friedman, J., & Rockoff, J. (2014c). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *American Economic Review*, *104*(9), 2593–2632. https://doi.org/10.1257/aer.104.9.2593

Cowan, J., Goldhaber, D., Hayes, K., & Theobald, R. (2016). Missing Elements in the Discussion of Teacher Shortages. *Educational Researcher*, *45*(8), 460–462. https://doi.org/10.3102/0013189X16679145

Cowan, J., Goldhaber, D., & Theobald, R. (2022). Performance Evaluations as a Measure of Teacher Effectiveness When Implementation Differs: Accounting for Variation across Classrooms, Schools, and Districts. *Journal of Research on Educational Effectiveness*. https://doi.org/https://doi.org/10.1080/19345747.2021.2018747

Dee, T. S., & Goldhaber, D. (2017). *Understanding and Addressing Teacher Shortages in the United States* (Issue April).

Dee, T., & Wyckoff, J. (2013). Incentives, Selection, and Teacher Performance: Evidence from IMPACT. *NBER Working Paper 19529*.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M.

(2010). *Impacts of Comprehensive Teacher Induction*. https://doi.org/10.1037/e599012011-001

Glazerman, S., & Seifullah, A. (2010). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report*. Mathematica Policy Research, Inc.

Goldhaber, D. (2007a). Everyone's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness? *Journal of Human Resources*, *42*(4), 765–794. http://www.appam.org/conferences/fall/madison2006/sessions/downloads/447510573.pdf

Goldhaber, D. (2007b). Teachers Matter, But Effective Teacher Quality Policies are Elusive: Hints from Research for Creating a More Productive Teacher Workforce. In H. F. Ladd & E. Fiske (Eds.), *Handbook of Research in Education Finance and Policy* (p. Section 10). Routledge.

Goldhaber, D. (2011). Licensure: Exploring the Value of this Gateway to the Teacher Workforce. In E. A. Hanushek, S. J. Machin, & L. Woessmann (Eds.), *Handbook of the Economics of Education* (Vol. 3, Issue 11, pp. 315–339). Elsevier B.V. https://doi.org/10.1016/B978-0-444-53429-3.00006-5

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2017). Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools. *Education Finance and Policy*, *12*(2), 197–223. https://doi.org/doi:10.1162/EDFP_a_00200

Goldhaber, D., Grout, C., Wolff, M., & Martinková, P. (2021). Evidence on the Dimensionality and Reliability of Professional References' Ratings of Teacher Applicants. *Economics of Education Review*, *83*(June). https://doi.org/10.1016/j.econedurev.2021.102130

Goldhaber, D., Quince, V., & Theobald, R. (2018). Has It Always Been This Way? Tracing the Evolution of Teacher Quality Gaps in U.S. Public Schools. *American Educational Research Journal*, *55*(1), 171–201. https://doi.org/10.3102/0002831217733445

Goldhaber, D., & Walch, J. (2012). Strategic Pay Reform : A Student Outcomes-Based Evaluation of Denver's ProComp Teacher Pay Initiative. *Economics of Education Review*, *31*(6), 1067–1083. http://www.sciencedirect.com/science/article/pii/S0272775712000751

Greenhaus, J. H., Parasuraman, S., & Wormley, W. M. (1990). Effects of Race on Organizational Experiences, Job Performance Evaluations, and Career Outcomes. *Academy of Management Journal*, *30*(1), 64–86. https://doi.org/https://doi-org.offcampus.lib.washington.edu/10.5465/256352

Gregory, R. G., & Borland, J. (1999). Recent developments in public sector labor markets. In *Handbook of Labor Economics* (pp. 3573–3630). Elsevier B.V.

Grissom, J. A., & Bartanen, B. (2022). Potential Race and Gender Biases in High-Stakes Teacher Observations. *Journal of Policy Analysis and Management*, *41*(1), 131–161. https://doi.org/10.1002/pam.22352

Heneman, H. G., & Judge, T. A. (2003). *Staffing Organizations* (4th ed.). McGraw-Hill/Mendota House.

Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non–Test Score Outcomes. *Journal of Political Economy*, *126*(5), 2072–2107.

Jacob, B. A., & Lefgren, L. (2008). Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education. *Journal of Labor Economics*, *26*(1), 101–136.

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools. *Journal of Public Economics*, *166*, 81–97. https://doi.org/https://doi.org/10.1016/j.jpubeco.2018.08.011

James, J., Kraft, M. A., & Papay, J. (2022). *Local Supply, Temporal Dynamics, and Unrealized Potential in Teacher Hiring* (No. 22-518; EdWorkingPaper). https://doi.org/10.26300/1yfe-gs84

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying Effective Classroom Practices Using Student Achievement Data. *Journal of Human Resources*, *46*(3), 587–613. https://doi.org/10.1353/jhr.2011.0010

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Larsen, B., Ziao, J., Kapor, A., & Yu, C. (2020). *The Effect of Occupational Licensing Stringency on the Teacher Quality Distribution* (No. 28158; NBER I Paper 28158). http://www.nber.org/papers/w28158

Leising, D., Erbs, J., & Fritz, U. (2010). The letter of recommendation effect in informant ratings of personality. *Journal of Personality and Social Psychology*, *98*(4), 668–682.

Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2009). Using the Standardized Letters of Recommendation in Selection: Results From a Multidimensional Rasch Model. *Educational and Psychological Measurement*, *69*(3), 475–492. https://doi.org/10.1177/0013164408322031

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., & Peng, A. (2011). *The Big Apple for Educators: New York City's Experiment with Schoolwide Performance Bonuses*. http://books.google.com/books?hl=en&lr=&id=k3vHxc0zY_MC&oi=fnd&pg=PP1&dq=marsh+pay+for+performance+2011&ots=1BkH5uttoZ&sig=PemMZsaQgXe0tBor4DwvPQ11 1lw#v=onepage&q&f=false

Mason, R. W., & Schroeder, M. P. (2014). The Predictive Validity of Teacher Candidate Letters of Reference. *Journal of Education and Learning*, *3*(3). https://doi.org/10.5539/jel.v3n3p67

McCarthy, J. M., & Goffin, R. D. (2001). Improving the Validity of Letters of Recommendation: An Investigation of Three Standardized Reference Forms. *Military Psychology*, *13*(4), 199–222. https://doi.org/10.1207/S15327876MP1304_2

National Council on Teacher Quality. (2014). *NCTQ Teacher Contract Database: NCTQ District Policy*.

Oyer, P., & Schaefer, S. (2011). Personnel Economics: Hiring and Incentives. In D. Car & O. Ashenfelter (Eds.), *Handbook of Labor Economics* (Volume 4, Vol. 4, Issue PART B, pp. 1769–1823). Elsevier B.V. https://doi.org/10.1016/S0169-7218(11)02418-X

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, Schools, and Academic Achievement.

Econometrica, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, *6*(1), 43–74. http://www.mitpressjournals.org.offcampus.lib.washington.edu/doi/pdf/10.1162/EDFP_a_00022

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2018). *Machine Learning and Applicant Work History*.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/https://doi.org/10.1037/apl0000405

Salgado, J. F. (2001). Personnel Selection Methods. In I. T. Robertson & C. L. Cooper (Eds.), *Personnel Psychology and Human Resource Management: A Reader for Students and Practitioners* (pp. 1–54). John Wiley & Sons, LTD.

Shaw, K. L., & Lazear, E. P. (2007). Personnel Economics: The Economist's View of Human Resources. *Journal of Economic Perspectives*, *21*(4), 91–114.

Society for Industrial and Organizational Psychology. (2014). *Employment Testing*.

Springer, M. G., Ballou, D., Hamilton, L. S., Le, V.-N., Lockwood, J., McCaffrey, D. F., Pepper, M., & Stecher, B. M. (2010). *Teacher Pay For Performance: Experimental Evidence from the Project on Incentives in Teaching: Vol. Project on*. RAND Corporation. http://www.rand.org/pubs/reprints/RP1416.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118.

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

Vergara vs. State of California Tentative Decision, (2014).

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting With Teacher Professional Development: Motives and Methods. *Educational Researcher*, *37*(8), 469–479. https://doi.org/10.3102/0013189X08327154

Tables

Table 1. Descriptive Statistics

	All	Employed Subsequent to Application				Not
		Spokane Public Schools			Other	
	Applicants	New	Transfer	Stay	District	Employed
	(1)	(2)	(3)	(4)	(5)	(6)
Observations						
Ratings	11,678	2,081	365	589	2,659	5,984
Unique Applicants	3,588	576	119	219	806	1,868
Rating on Overall Criterion						
No Basis for Judgment	1%	0%	0%	0%	1%	1%
Below Average	1%	1%	1%	1%	1%	1%
Average	6%	5%	6%	5%	5%	8%
Very Good	15%	13%	13%	16%	14%	16%
Excellent	23%	21%	22%	20%	25%	23%
Outstanding (top 5%)	34%	36%	36%	38%	34%	32%
Among Best (top 1%)	20%	24%	22%	21%	20%	19%
GRM Measure*	0.00	0.11	0.05	0.08	0.02	-0.07
Rater Type						
Principal	35%	37%	48%	49%	38%	31%
Dept. Chair/Instr. Coach	5%	5%	7%	8%	4%	4%
Colleague	26%	25%	29%	29%	25%	26%
Cooperating Teacher	12%	14%	5%	4%	14%	12%
University Supervisor	9%	10%	4%	4%	10%	9%
Other	13%	10%	7%	6%	9%	16%
Applicant Experience**						
Novice	30%	29%	0%	0%	27%	37%
1 year	11%	19%	22%	13%	18%	5%
2 to 5 years	16%	24%	30%	30%	22%	10%
6+ years	23%	28%	48%	58%	33%	11%
Coverage of Outcome Measures						
Performance Evals.	23%	91%	61%	50%	N/A	N/A
Value-Added in Math	7%	14%	11%	9%	15%	N/A
Value-Added in Reading	7%	14%	9%	8%	18%	N/A

Notes: Individuals who apply for positions in multiple hiring years are treated as distinct applicants. An applicant is considered "Employed" if they are subsequently observed in a classroom teaching position according to the S-275 personnel records maintained by Washington State.

GRM* refers to the summative reference ratings measure described in **Section 3.2. Because it utilizes information from each rating criterion, it is not calculated in cases where one or more criteria is rated as "No Basis for Judgment" (1,253 observations in column (1)).

**Applicant experience is based on the level of experience reported in the S-275 personnel records and is available for a subset of applicants in columns (1) (9,407 observations) and (6) (3,728 observations).

Table 2. Predicting Performance Evaluations

	(1)	(2)	(3)	(4)
GRM	0.148***	0.129***		
	(0.028)	(0.029)		
Overall Criterion				
Average/Below Average			-0.205	-0.255
			(0.151)	(0.159)
Very Good			Ref.	Ref.
			-	-
Excellent (top 10%)			0.117	0.106
			(0.073)	(0.070)
Outstanding (top 5%)			0.228***	0.212**
			(0.068)	(0.065)
Among the best (top 1%)			0.402***	0.336***
			(0.079)	(0.077)
Teacher Controls		X		X
R^2	0.022	0.057	0.024	0.061
Observations	2,439	2,439	2,439	2,439
Clusters/Teachers	757	757	757	757

Notes: GRM is the standardized summative reference ratings measure described in Section 3.2. Teacher controls include indicators for whether they are female, white, or hold an advanced degree, and a categorical control for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the teacher level.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Heterogeneity in Predicting Performance Evaluations by Applicant Type

	(1)	(2)	(3)	(4)
Panel A – Novice Applicants				
GRM	0.020	0.019		
	(0.051)	(0.051)		
Overall Criterion				
Average/Below Average			0.012	-0.037
			(0.218)	(0.219)
Very Good			Ref.	Ref.
			-	-
Excellent (top 10%)			-0.041	-0.039
			(0.133)	(0.133)
Outstanding (top 5%)			0.152	0.144
			(0.118)	(0.119)
Among the best (top 1%)			0.150	0.133
			(0.136)	(0.131)
Teacher Controls		X		X
R^2	0.002	0.021	0.011	0.029
Observations	598	598	598	598
Clusters/Teachers	177	177	177	177
Panel B – Experienced Applicants				
GRM	0.171***	0.166***		
	(0.032)	(0.032)		
Overall Criterion				
Average/Below Average			-0.265	-0.308
			(0.179)	(0.179)
Very Good			Ref.	Ref.
			-	-
Excellent (top 10%)			0.171*	0.155
			(0.086)	(0.083)
Outstanding (top 5%)			0.258**	0.244**
			(0.082)	(0.079)
Among the best (top 1%)			0.463***	0.432***
			(0.094)	(0.091)
Teacher Controls		X		X
R^2	0.028	0.043	0.031	0.047
Observations	1,841	1,841	1,841	1,841
Clusters/Teachers	580	580	580	580

Notes: Novice applicants are those who do not report prior teaching experience in their application profiles. *GRM* is the standardized summative reference ratings measure described in Section 3.2. Teacher controls include indicators for whether they are female, white, or hold an advanced degree. Standard errors are clustered at the teacher level.* p < 0.05, ** p < 0.01, *** p < 0.001

Table 4. Heterogeneity in Predicting Performance Evaluations by Rater Type

	(1)	(2)	(3)	(4)
Rater Type				
Principal	0.128	0.131	0.256**	0.242**
	(0.090)	(0.082)	(0.095)	(0.086)
Instr. Coach/Dept. Chair	0.149	0.188	0.257*	0.271*
	(0.118)	(0.122)	(0.122)	(0.126)
Colleague	Ref.	Ref.	Ref.	Ref.
	-	-	-	-
Cooperating Teacher	-0.158*	-0.031	-0.077	0.004
	(0.079)	(0.067)	(0.083)	(0.069)
University Supervisor	-0.149	-0.029	-0.076	0.002
	(0.109)	(0.084)	(0.113)	(0.086)
Other	-0.087	0.005	-0.012	0.050
	(0.114)	(0.109)	(0.120)	(0.114)
GRM				
GRM*Principal			0.193***	0.181***
			(0.044)	(0.046)
GRM*Instr. Coach/Dept. Chr.			0.235***	0.195**
			(0.070)	(0.072)
GRM*Colleague			0.266***	0.227***
			(0.054)	(0.051)
GRM*Cooperating Teacher			0.019	0.016
			(0.052)	(0.052)
GRM*University Supervisor			0.102	0.078
			(0.069)	(0.069)
GRM*Other			0.039	0.049
			(0.080)	(0.080)
Teacher Controls		X		X
R^2	0.014	0.046	0.049	0.073
Observations	2,439	2,439	2,439	2,439
Clusters/Teachers	757	757	757	757

GRM is the standardized summative reference ratings measure described in Section 3.2. Teacher controls include indicators for whether they are female, white, or hold an advanced degree, and a categorical control for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the teacher level.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. Predicting Performance Evaluations Using Individual Evaluation Criteria

	(1)	(2)	(3)	(4)	(5)	(6)
Rating						
Avg./Below Avg.	-0.126	-0.241	-0.133	-0.389*	-0.220	-0.013
	(0.120)	(0.129)	(0.185)	(0.167)	(0.151)	(0.110)
Very Good	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	-	-	-	-	-	-
Excellent (top 10%)	0.077	0.046	0.178**	-0.037	0.064	0.134
	(0.063)	(0.070)	(0.068)	(0.092)	(0.067)	(0.078)
Outstanding (top 5%)	0.142*	0.068	0.241***	0.067	0.136*	0.245**
	(0.068)	(0.068)	(0.068)	(0.085)	(0.063)	(0.078)
Among best (top 1%)	0.301***	0.353***	0.326***	0.211*	0.354***	0.452***
	(0.079)	(0.077)	(0.080)	(0.086)	(0.075)	(0.091)
Evaluation Criterion	Challenges Students	Classroom Mgmt.	Working w/ Diverse Groups of Students	Interpersonal Skills	Student Engagement	Instructional Skills
R^2	0.013	0.026	0.014	0.019	0.023	0.024
Observations	2,439	2,439	2,439	2,439	2,439	2,439
Clusters/Teachers	757	757	757	757	757	757

Notes: Each column is a separate regression model analyzing reference ratings of a different evaluation criterion. Standard errors are clustered at the teacher level.* p < 0.05, ** p < 0.01, *** p < 0.001

Table 6. Predicting Performance Evaluations Using Strongest/Weakest Ratings

	Strongest (1)	Strongest (2)	Weakest (3)	Weakest (4)
Strongest/Weakest Competency				
Challenges Students	-0.031	-0.047	-0.039	-0.049
	(0.095)	(0.092)	(0.097)	(0.093)
Classroom Management	0.042	0.007	-0.000	-0.024
	(0.078)	(0.073)	(0.078)	(0.073)
Instructional Skills	0.030	-0.028	-0.041	-0.087
	(0.081)	(0.077)	(0.084)	(0.081)
Interpersonal Skills	Ref.	Ref.	Ref.	Ref.
	-	-	-	-
Student Engagement	0.037	0.010	-0.015	-0.030
	(0.071)	(0.067)	(0.073)	(0.069)
Working with Diverse Groups of Students	-0.117	-0.151	-0.148	-0.172*
	(0.094)	(0.092)	(0.088)	(0.085)
Teacher Controls		X		X
R^2	0.004	0.044	0.029	0.064
Observations	2,439	2,439	2,439	2,439
Clusters/Teachers	757	757	757	757

Notes: Teacher controls include indicators for whether they are female, white, or hold an advanced degree, and a categorical control for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the teacher level.* p < 0.05, ** p < 0.01, *** p < 0.001

Table 7. Predicting Teacher Value-Added

	Math				Reading			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
GRM	0.020*	0.025**			0.013	0.014		
	(0.009)	(0.009)			(0.010)	(0.010)		
Overall Criterion								
Avg./Below Avg.			-0.030	-0.034			0.006	0.009
			(0.034)	(0.034)			(0.034)	(0.035)
Very Good			Ref.	Ref.			Ref.	Ref.
			-	-			-	-
Excellent			0.060*	0.059*			0.035	0.034
			(0.026)	(0.025)			(0.027)	(0.026)
Outstanding			0.041	0.044*			0.044	0.045
			(0.022)	(0.022)			(0.024)	(0.024)
Among best			0.061*	0.070**			0.030	0.033
			(0.028)	(0.027)			(0.033)	(0.033)
Teacher Controls		X		X		X		X
R^2	0.009	0.036	0.017	0.042	0.003	0.007	0.005	0.008
Observations	793	793	793	793	804	804	804	804
Clusters/Teachers	268	268	268	268	268	268	268	268

Notes: GRM is the standardized summative reference ratings measure described in Section 3.2. Teacher controls include a categorical variable for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the applicant level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8. Heterogeneity in Predicting Teacher Value-Added by Applicant Type

	Math (1)	Math (2)	Reading (3)	Reading (4)
Panel A – Novice Applicants				
GRM	0.007		0.005	
	(0.015)		(0.016)	
Overall Criterion				
Avg./Below Avg.		0.022		0.053
		(0.055)		(0.046)
Very Good		Ref.		Ref.
		-		-
Excellent		0.049		0.017
		(0.037)		(0.039)
Outstanding		0.034		0.051
		(0.037)		(0.033)
Among Best		0.058		0.016
		(0.054)		(0.052)
R^2	0.001	0.009	0.001	0.013
Observations	248	248	236	236
Clusters/Teachers	81	81	77	77
Panel B – Experienced Applicants				
GRM	0.025*		0.016	
	(0.011)		(0.013)	
Overall Criterion				
Avg./Below Avg.		-0.046		-0.016
		(0.042)		(0.045)
Very Good		Ref.		Ref.
		-		-
Excellent		0.068		0.044
		(0.036)		(0.035)
Outstanding		0.045		0.042
		(0.028)		(0.032)
Among best		0.064		0.033
		(0.033)		(0.041)
R^2	0.015	0.023	0.005	0.006
Observations	545	545	568	568
Clusters/Teachers	187	187	191	191

Notes: Novice applicants are those who do not report prior teaching experience in their application profiles. *GRM* is the standardized summative reference ratings measure described in Section 3.2. Standard errors are clustered at the applicant level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9. Heterogeneity in Predicting Teacher Value-Added by Rater Type

	Math				Reading			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Rater Type								
Principal	-0.067**	-0.071***	-0.066**	-0.068**	-0.064**	-0.065**	-0.066*	-0.065*
	(0.021)	(0.020)	(0.023)	(0.022)	(0.025)	(0.025)	(0.026)	(0.026)
Coach/Dept. Chr.	-0.111*	-0.122**	-0.102*	-0.108*	-0.152***	-0.155***	-0.146***	-0.148***
	(0.043)	(0.043)	(0.045)	(0.044)	(0.042)	(0.042)	(0.043)	(0.044)
Colleague	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	-	-	-	-	-	-	-	-
Coop. Teacher	-0.058*	-0.066**	-0.065*	-0.074**	-0.068*	-0.069*	-0.071*	-0.074*
	(0.028)	(0.024)	(0.029)	(0.025)	(0.030)	(0.029)	(0.030)	(0.030)
University Sup.	-0.054	-0.062**	-0.056	-0.065**	-0.054	-0.055	-0.051	-0.054
	(0.030)	(0.023)	(0.031)	(0.024)	(0.030)	(0.030)	(0.030)	(0.030)
Other	-0.065*	-0.073*	-0.068*	-0.078*	-0.090**	-0.094*	-0.085*	-0.090*
	(0.031)	(0.032)	(0.031)	(0.031)	(0.034)	(0.036)	(0.036)	(0.038)
GRM								
GRM*Principal			0.038**	0.043**			0.019	0.022
			(0.014)	(0.014)			(0.014)	(0.014)
GRM*Coach/Chair			0.053	0.069*			0.055*	0.057*
			(0.031)	(0.029)			(0.028)	(0.028)
GRM*Colleague			-0.013	-0.005			-0.008	-0.003
			(0.019)	(0.019)			(0.020)	(0.020)
GRM*Coop. Tchr.			-0.017	-0.013			-0.020	-0.019
			(0.022)	(0.021)			(0.022)	(0.022)
GRM*Univ. Sup.			0.013	0.015			0.029	0.030
			(0.025)	(0.024)			(0.024)	(0.023)
GRM*Other			-0.003	0.005			-0.025	-0.022
			(0.032)	(0.031)			(0.043)	(0.042)
Teacher Controls		X		X		X		X
R^2	0.021	0.046	0.040	0.070	0.032	0.035	0.043	0.047
Observations	793	793	793	793	804	804	804	804
Clusters/Teachers	268	268	268	268	268	268	268	268

Notes: GRM is the standardized summative reference ratings measure described in Section 3.2. Teacher controls include a categorical variable for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the applicant level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10. Predicting Teacher Value-Added Using Individual Evaluation Criteria

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A - Math						
Rating						
Avg./Below Avg.	0.031	0.026	0.011	0.055	0.004	0.000
	(0.033)	(0.030)	(0.047)	(0.043)	(0.030)	(0.034)
Very Good	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	-	-	-	-	-	-
Excellent (top 10%)	0.042	0.061*	0.044	0.008	0.044	0.048*
	(0.026)	(0.026)	(0.024)	(0.029)	(0.025)	(0.022)
Outstanding (top 5%)	0.040	0.059**	0.053**	0.022	0.068**	0.058**
	(0.021)	(0.018)	(0.021)	(0.029)	(0.024)	(0.021)
Among best (top 1%)	0.056*	0.068**	0.039	0.024	0.053*	0.076**
	(0.028)	(0.026)	(0.025)	(0.032)	(0.026)	(0.028)
Evaluation Criterion	Challenges Students	Classroom Mgmt.	Working w/ Diverse Groups of Students	Interpersonal Skills	Student Engagement	Instructional Skills
R^2	0.008	0.014	0.009	0.003	0.014	0.017
Observations	793	793	793	793	793	793
Clusters/Teachers	268	268	268	268	268	268
Panel B - Reading						
Rating						
Avg./Below Avg.	0.011	0.065	0.028	0.017	-0.025	0.005
	(0.031)	(0.033)	(0.045)	(0.055)	(0.037)	(0.036)
Very Good	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	-	-	-	-	-	-
Excellent (top 10%)	0.028	0.052	0.049	0.005	0.037	0.035
	(0.026)	(0.027)	(0.027)	(0.033)	(0.025)	(0.023)
Outstanding (top 5%)	0.038	0.070**	0.053*	-0.005	0.057*	0.056*
	(0.021)	(0.021)	(0.024)	(0.031)	(0.026)	(0.023)
Among best (top 1%)	0.034	0.067*	0.032	-0.006	0.038	0.044
	(0.029)	(0.028)	(0.030)	(0.036)	(0.030)	(0.032)
Evaluation Criterion	Challenges Students	Classroom Mgmt.	Working w/ Diverse Groups of Students	Interpersonal Skills	Student Engagement	Instructional Skills
R^2	0.004	0.014	0.007	0.001	0.011	0.009
Observations	804	804	804	804	804	804
Clusters/Teachers	268	268	268	268	268	268

Notes: Standard errors are clustered at the teacher level.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11. Predicting Teacher Value-Added Using Strongest/Weakest Ratings

	Strongest (1)	Strongest (2)	Weakest (3)	Weakest (4)
Panel A – Value-Added in Math				
Strongest/Weakest Competency				
Challenges Students	0.042	0.047	-0.031	-0.035
	(0.030)	(0.029)	(0.023)	(0.023)
Classroom Management	0.056*	0.067*	-0.036	-0.040
	(0.028)	(0.029)	(0.024)	(0.025)
Instructional Skills	0.040	0.049*	0.003	-0.011
	(0.022)	(0.022)	(0.032)	(0.033)
Interpersonal Skills	Ref.	Ref.	Ref.	Ref.
	-	-	-	-
Student Engagement	0.006	0.010	-0.026	-0.031
	(0.026)	(0.026)	(0.032)	(0.032)
Working with Diverse Groups of Students	0.005	0.016	-0.002	0.000
	(0.025)	(0.025)	(0.027)	(0.027)
Teacher Controls		X		X
R^2	0.011	0.036	0.007	0.030
Observations	793	793	793	793
Clusters/Teachers	268	268	268	268
Panel B – Value-Added in Reading				
Strongest/Weakest Competency				
Challenges Students	0.021	0.022	-0.052	-0.053
	(0.044)	(0.043)	(0.029)	(0.029)
Classroom Management	0.023	0.023	-0.044	-0.043
	(0.031)	(0.031)	(0.028)	(0.029)
Instructional Skills	-0.024	-0.024	-0.022	-0.024
	(0.024)	(0.025)	(0.034)	(0.033)
Interpersonal Skills	Ref.	Ref.	Ref.	Ref.
	-	-	-	-
Student Engagement	-0.017	-0.018	-0.073	-0.073
	(0.027)	(0.027)	(0.038)	(0.038)
Working with Diverse Groups of Students	-0.038	-0.037	-0.010	-0.008
	(0.024)	(0.024)	(0.029)	(0.029)
Teacher Controls		X		X
R^2	0.010	0.013	0.013	0.015
Observations	804	804	804	804
Clusters/Teachers	268	268	268	268

Notes: Teacher controls include a categorical control for experience (0, 1, 2-5, and 6+ years). Standard errors are clustered at the teacher level.* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12. Heckman Selection Models – Performance Evaluations

	Selection into the Sample	Performance Outcome (Corrected)	Performance Outcome (Uncorrected)
	(1)	(2)	(3)
GRM	0.154***	0.222***	0.168***
	(0.024)	(0.077)	(0.031)
Excluded Variables			
Quantity of Competition	-0.006**		
	(0.002)		
Quality of Competition	-0.046		
	(0.120)		
Selection Correction			
Inverse Mills Ratio (λ)		0.486	
		(0.612)	
Observations	7,148	2,187	2,187

Notes: The model is estimated using Stata's *heckman* command as a two-step model with bootstrapped standard errors (500 replications) clustered at the applicant level. The first stage of the model is presented in column (1) with selection into the sample defined as an applicant being hired for a position to which they applied *and* having an observed performance outcome. The selection-corrected performance outcome model is presented in column (2). An uncorrected performance outcome model is presented in column (3). *GRM* is the standardized summative reference ratings measure described in Section 3.2. Each model includes controls for ethnicity, whether the applicant holds an advanced degree, a categorical control for experience (0, 1, 2-5, and 6+ years), and rater type. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

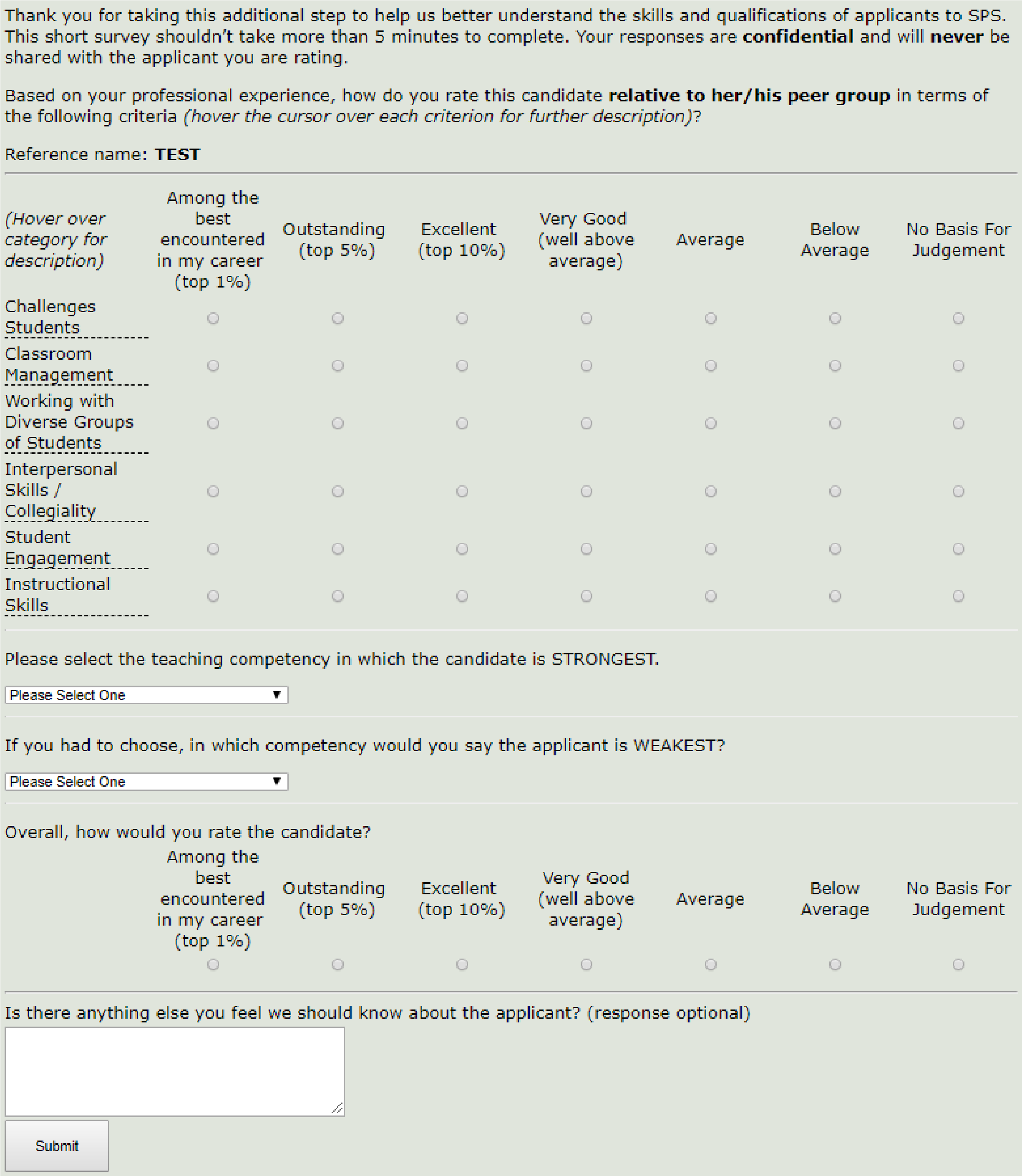
Table 13. Heckman Selection Models – Teacher Value-Added

	Selection into the Sample	Performance Outcome (Corrected)	Performance Outcome (Uncorrected)
	(1)	(2)	(3)
Panel A – Math			
GRM	0.042	0.022	0.022
	(0.038)	(0.013)	(0.012)
Excluded Variables			
Quantity of Competition	-0.012**		
	(0.004)		
Quality of Competition	0.603		
	(0.341)		
Selection Correction			
Inverse Mills Ratio (λ)		0.105	
		(0.114)	
Observations	2,857	263	263
Clusters	903	85	85
Panel B – Reading			
GRM	0.071*	0.002	-0.004
	(0.035)	(0.019)	(0.013)
Excluded Variables			
Quantity of Competition	-0.008		
	(0.004)		
Quality of Competition	0.552*		
	(0.253)		
Selection Correction			
Inverse Mills Ratio (λ)		0.135	
		(0.1)	
Observations	3,338	283	283
Clusters	1,054	94	94

Notes: The model is estimated using Stata's *heckman* command as a two-step model with bootstrapped standard errors (500 replications) clustered at the applicant level. The first stage of the model is presented in column (1) with selection into the sample defined as an applicant being hired for a position to which they applied *and* having an observed performance outcome. The selection-corrected performance outcome model is presented in column (2). An uncorrected performance outcome model is presented in column (3). Each model includes controls for ethnicity, whether the applicant holds an advanced degree, a categorical control for experience (0, 1, 2-5, and 6+ years), endorsement area indicators, and rater type. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figures

Figure 1. Professional Reference Survey Form



Thank you for taking this additional step to help us better understand the skills and qualifications of applicants to SPS. This short survey shouldn't take more than 5 minutes to complete. Your responses are **confidential** and will **never** be shared with the applicant you are rating.

Based on your professional experience, how do you rate this candidate **relative to her/his peer group** in terms of the following criteria *(hover the cursor over each criterion for further description)*?

Reference name: **TEST**

(Hover over category for description)	Among the best encountered in my career (top 1%)	Outstanding (top 5%)	Excellent (top 10%)	Very Good (well above average)	Average	Below Average	No Basis For Judgement
Challenges Students	○	○	○	○	○	○	○
Classroom Management	○	○	○	○	○	○	○
Working with Diverse Groups of Students	○	○	○	○	○	○	○
Interpersonal Skills / Collegiality	○	○	○	○	○	○	○
Student Engagement	○	○	○	○	○	○	○
Instructional Skills	○	○	○	○	○	○	○

Please select the teaching competency in which the candidate is STRONGEST.

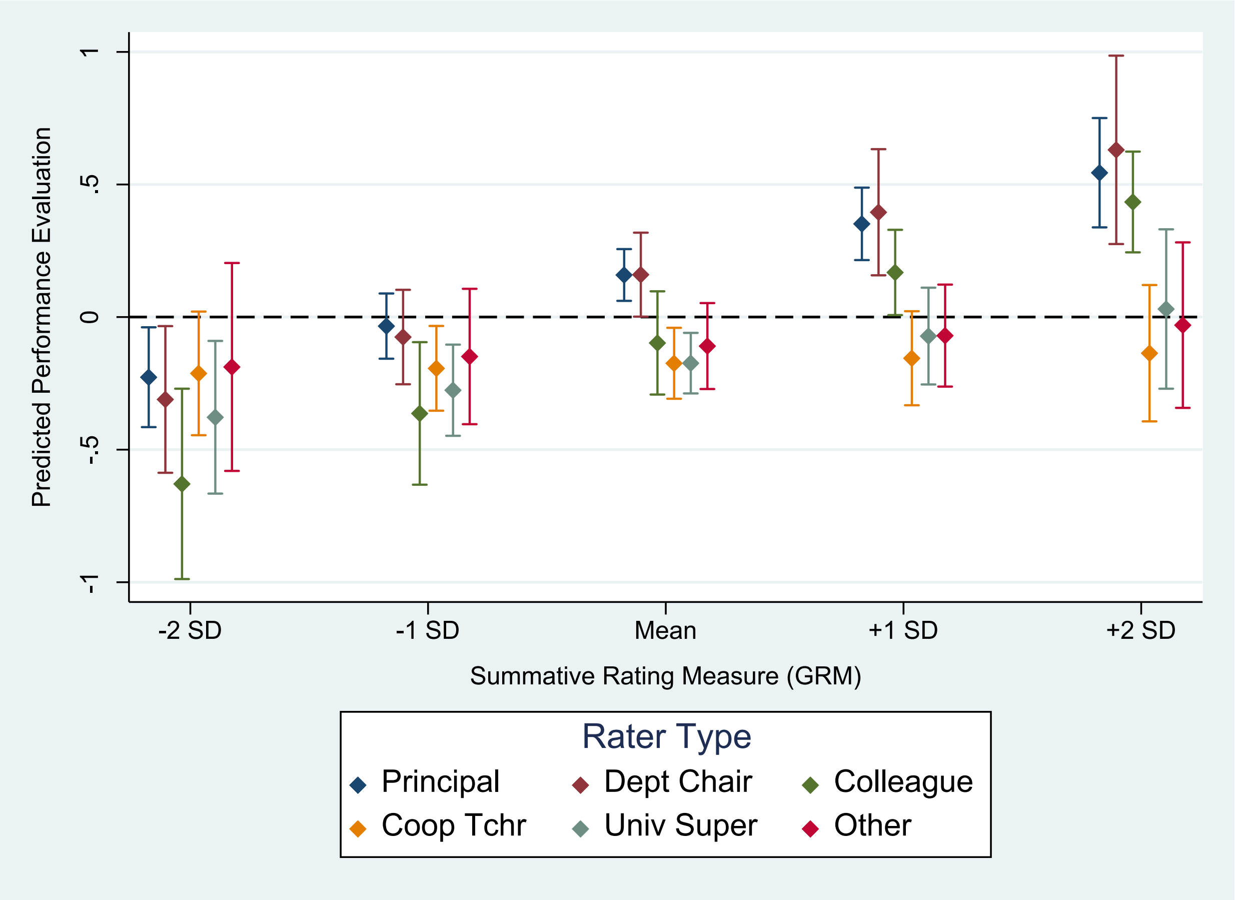
If you had to choose, in which competency would you say the applicant is WEAKEST?

Overall, how would you rate the candidate?

Among the best encountered in my career (top 1%)	Outstanding (top 5%)	Excellent (top 10%)	Very Good (well above average)	Average	Below Average	No Basis For Judgement
○	○	○	○	○	○	○

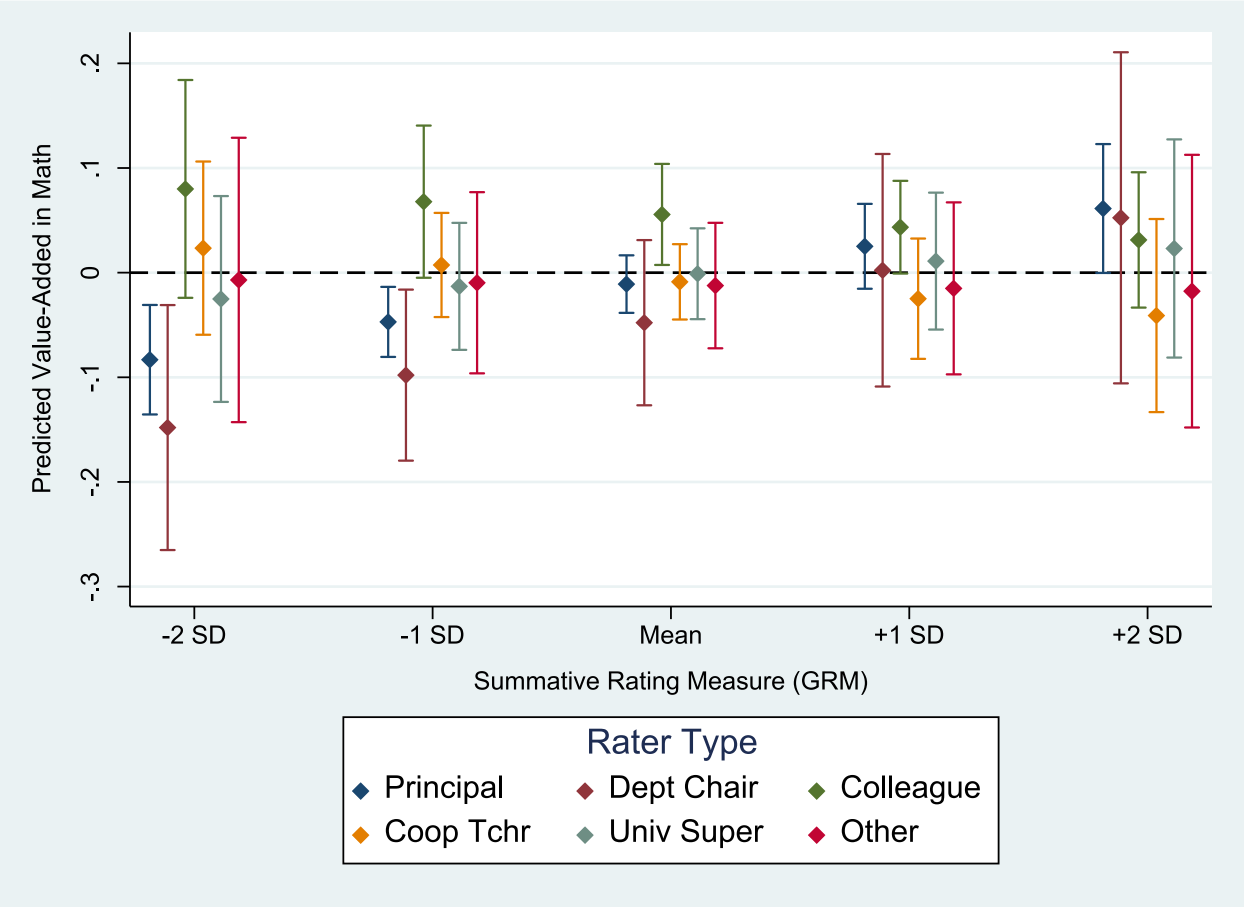
Is there anything else you feel we should know about the applicant? (response optional)

Figure 2. Predicted Performance Evaluations by Summative Rating Level and Rater Type



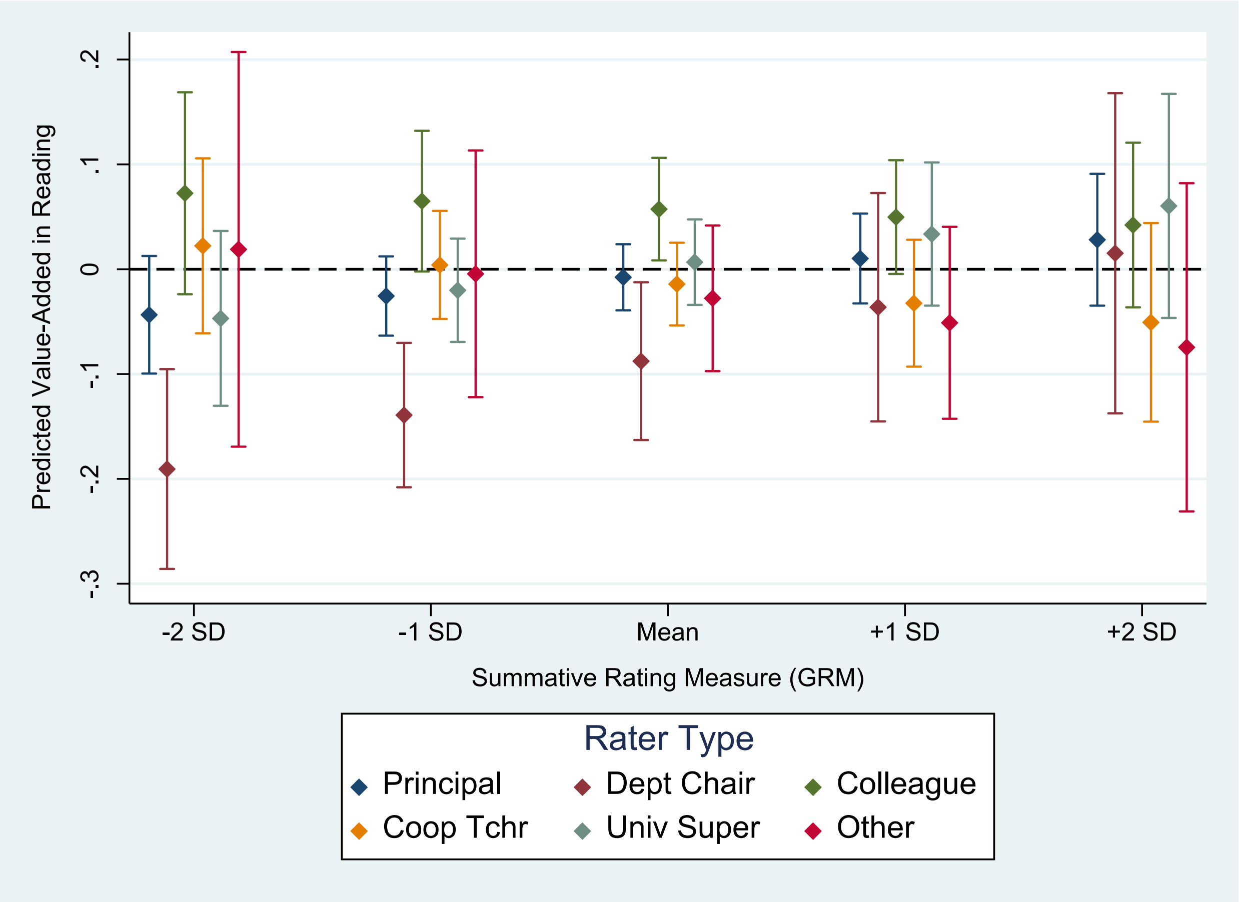
Notes: Predicted values and 95% confidence intervals are generated based on the regression output represented in column (3) of **Table 4**.

Figure 3. Predicted Value Added to Math by Summative Rating Level and Rater Type



Notes: Predicted values and 95% confidence intervals are generated based on the regression output represented in column (3) of **Table 9**.

Figure 4. Predicted Value Added to Reading by Summative Rating Level and Rater Type



Notes: Predicted values and 95% confidence intervals are generated based on the regression output represented in column (3) of **Table 9**.

Appendix A. Supplemental Figures

Figure A1: Distribution of “Overall” Rating by Rater Type

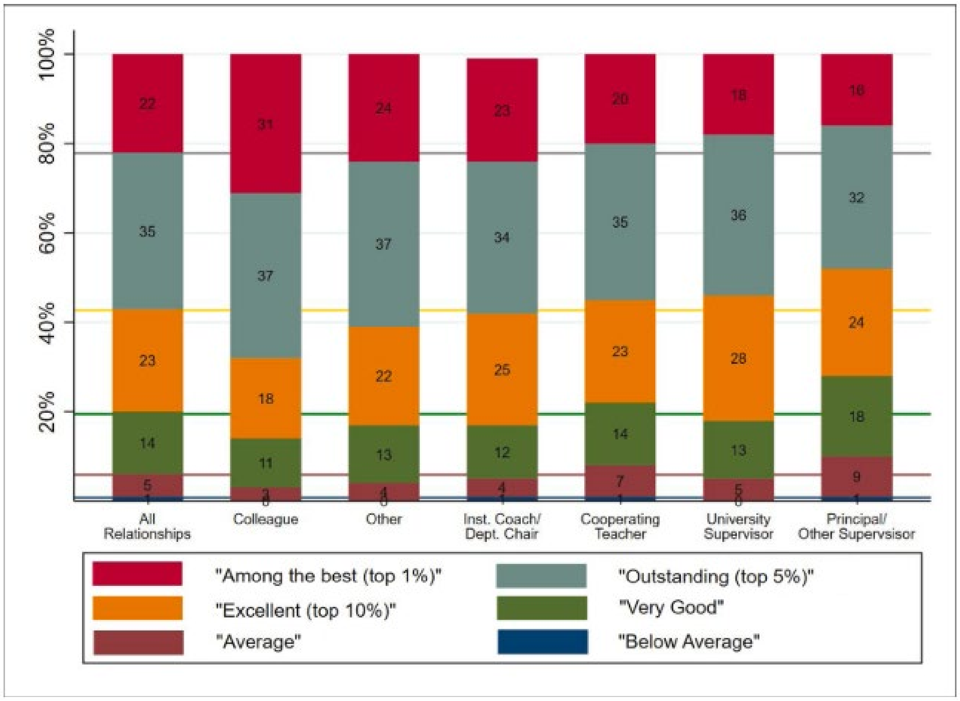


Fig. 2. Distribution of Ratings on “Overall” Criterion by Rater Type Distribution of ratings by applicant-rater relationship type (N = 10,763).

Source: Goldhaber et al. (2021), page 6.

Appendix B. Supplemental Tables

Table B1. Description of Criteria for References' Ratings of Applicants

Criterion	Description
Student Engagement	• Lessons interest and engage students • Teacher is effective at relating to students
Instructional Skills	• Establishes clear learning objectives and monitors progress • Teacher utilizes multiple approaches to reach different types of students • Ability to adapt curriculum and teaching style to new state and federal requirements
Classroom Management	• Develops routines and procedures to increase learning. • Is effective at maintaining control of the classroom (this may not mean quiet and orderly, but planned and directed) • Students in class treat one another with respect
Working with Diverse Groups of Students	• Is effective at encouraging and relating to students from disadvantaged backgrounds
Interpersonal Skills	• Develops and maintains effective working relationship with colleagues • Contributes to establishing a positive classroom and school environment • Interactions with parents are productive
Challenges Students	• Sets high expectations and holds students accountable