CEDR Working Paper | No. 08262024-1

CEDR
Center for Education
Data & Research

Improving Hiring Decisions: Experimental Evidence on the Value of Reference Information About Teacher Applicants

Dan Goldhaber
University of Washington
American Institutes for Research

Cyrus Grout
University of Washington

The Center for Education Data & Research
University of Washington Bothell
3610 Albion Pl. N.   Ste. A   Seattle, WA 98103
(206) 547-5585   cedr@u.washington.edu
www.cedr.us

Acknowledgements: 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 study's funder. We thank Spokane Public Schools for partnering with us on this work and James Cowan, Michael DeArmond, Edie Harding, Heather Hill, Susana Loeb, and Jonah Rockoff for helpful feedback and suggestions on an earlier draft of this paper. Correspondence regarding this article should be addressed to Dan Goldhaber, Center for Education Data and Research, University of Washington, 3610 Albion Pl. N, Suite A, Seattle, WA 98103. Email: dgoldhab@uw.edu • www.cedr.us

Suggested citation:
Goldhaber, D.; Grout, C. (2024). Improving Hiring Decisions: Experimental Evidence on the Value of Reference Information About Teacher Applicants. CEDR Working Paper 08262024-1. University of Washington, Seattle, WA.

Abstract: Professional references are widely used in hiring decisions, yet their effectiveness remains largely understudied. We analyze structured ratings collected from the professional references of teacher applicants and conduct an experiment to see whether the ratings influence hiring managers' assessments of applicants and hiring decisions. We find little evidence that providing reference ratings to hiring managers influences their evaluations of candidates or hiring choices in productive ways. Importantly, we also find that reference ratings are predictive of future job performance. The result is a paradox: reference ratings offer potentially low-cost, high-value information, but hiring managers do not appear to make productive use of them.

Contents

Contents i

1. Introduction 1

2. Background on Hiring Practices and Indicators of Applicant Productivity 3

3. Study Setting and Information Experiment 5

4. Empirical Approach 17

5. Results 22

6. Discussion of Scope for Change 29

7. Conclusions 33

References 36

Tables and Figures 40

Appendix: Supplemental Tables and Figures 49

1. Introduction

A well-developed literature in personnel economics emphasizes the challenges of making good hiring decisions (e.g., Bloom and Van Reenen, 2011; Bilan et al., 2020; Heneman and Judge, 2003; Lazear and Shaw, 2007). Determining who will be effective is difficult; asymmetric information abounds, as job seekers have more information about their own skills and motivation than hiring managers (Jovanovic, 1979; Montgomery, 1991). In an ideal world, hiring managers would gather low-cost, high-value information about job candidates to make better hiring decisions. But research currently offers little guidance on what that might entail (Oyer & Shaefer, 2011). In this study, we examine the issue in the context of one of the country's largest professions: public school teachers. We report on the extent to which low-cost, high-value information collected from professional references influences hiring manager assessments of applicants and hiring decisions.

We ask professional references of teacher applicants to Spokane Public Schools (henceforth "Spokane") to provide *categorical ratings* of applicants (we refer to these as "reference ratings") in addition to the letters of recommendation they already provide. We conduct an experiment in which we provide the reference ratings to hiring managers for a random subset of job applicants, obtaining causal estimates of whether the reference ratings influence hiring managers' assessments of applicants and hiring choices. We fail to find evidence that providing the reference ratings information to hiring managers influences their evaluations of applicants or hiring decisions. Given the precision of the estimates, we can rule out changes smaller than 1 percentage point in hiring probability. This finding is consistent with Jacob et al. (2018), who study the teacher hiring process in Washington, DC, who also find that information available to hiring officials is predictive of in-service outcomes but is not utilized.

That the information from professional references does not influence hiring appears to be a lost opportunity as, consistent with prior evidence (Goldhaber et al., 2024), reference ratings are found to predict future job performance (they do not significantly predict teacher retention). Indeed, we provide novel evidence that reference ratings add to the predictive power of the information regularly collected and processed about applicants, suggesting the ratings indeed *add* information to the hiring process. The result is a paradox: reference ratings offer potentially low-cost, high-value information, but hiring managers do not appear to make productive use of them.

Our findings contribute to the broader field of personnel economics by providing evidence on the degree to which specific human resource management (HRM) practices are related to productivity and hiring (Bloom and Van Reenen, 2011). The profession we study is significant, given that teachers represent about 2% of the American workforce and over 1% of GDP in salary alone (Backes, 2024). At the same time, we contribute to the literature teacher hiring in education. The importance of hiring in K12 education is well-recognized, given the profound impacts teachers can have on short (Aaronson et al., 2007; Rivkin et al., 2005) and long-run (Backes et al., 2023; Chetty et al., 2014) student outcomes. Given evidence that low-performing teachers are unlikely to catch up with higher-performing peers (Atteberry et al., 2013) and that it is costly to remove ineffective teachers (National Council on Teacher Quality, 2014; Treu, 2014), poor hiring decisions in education carry extra weight.

Even in a tight labor market, some districts have a relatively large numbers of applicants for open positions (Goldhaber et al., 2017); in our sample there are roughly 6 unique applicants per open teaching slot. Often, however, districts fail to capitalize on their hiring opportunities, struggling to effectively screen and hire candidates. Typical hiring procedures in schools have often been characterized as rushed and information poor (Liu and Johnson, 2006; Papay and

Kraft, 2016), and a recent study found that the number of applicants for a teaching position was largely unrelated to the quality of the hire (James et al., 2023).[1] As James et al (2023) conclude, "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. 1040). Our findings underscore this point: improved screening requires both identifying valuable information *and* ensuring hiring managers are motivated to use it.

2. Background on Hiring Practices and Indicators of Applicant Productivity

In their handbook chapter examining the state of the personnel economics literature on hiring, Oyer and Schaefer (2011) frame the economic problem in hiring as a matching problem, one "with costly search and bilateral asymmetric information. Job seekers have varying levels of aptitude, skill, and motivation, and firms have varying needs for these attributes" (p. 1784). The problem is that gathering information to make a good person-job match is costly for both employers and jobseekers. But despite being well conceptualized, there is little empirical evidence on how different hiring practices impact organizational performance. As Oyer and Schaefer (2011) note, several studies link the adoption of bundles of human resource practices to greater firm productivity (Bresnahan, 2002; Bloom and Van Reenen, 2007; Ichniowski et al., 1997), but they do not separately identify the effects of hiring practices. Huang and Cappelli (2010), present correlational evidence that employers who report screening candidates more intensively for attributes related to work ethic exhibit higher employee productivity and lower

[1] Studies based on observations of teachers in the workforce (but not applicants or job offers) find mixed evidence about whether school districts hire better teachers in weak labor markets (Nagler et al., 2020; Rucinski, 2023). Staiger and Rockoff (2010) also describe evidence that districts struggle to discern teacher effectiveness during the hiring process. In the late 1990's, the state of California began providing incentives to keep K-3 class sizes at a maximum of 20 children, spurring a dramatic increase in the hiring of elementary teachers in Los Angeles Unified School District. Had the district been effective at discerning teacher effectiveness, one would expect the average effectiveness of teachers hired during this hiring bubble to be lower than that of previously hired teachers, but the value-added of teachers hired during the bubble was no worse than that of previously hired teachers.

rates of involuntary turnover. But like much of the research in this area, the authors do not establish a causal relationship or compare different hiring practices. And so, employers are left grappling with the costs of obtaining information about prospective hires and uncertainty about what information might lead to better outcomes.

Hiring managers are not entirely without guidance, however. Research in organizational psychology suggests, for example, that certain applicant screening procedures can provide valuable information about job applicants. In a meta-analysis of meta-analyses on the operational validity of different personnel selection methods, Sackett et al. (2021) find that measures of applicant quality from structured interviews, job knowledge tests, keyed biographical data, work sample tests, and cognitive ability tests exhibited the highest levels of validity. However, the research in this area also has limitations, especially regarding outcomes. Many studies fail to directly link screening measures to employee performance, and those that do often rely on simple bivariate correlations between applicant measures and subsequent performance (e.g., Judge et al., 2007).

A smaller but growing literature on public school hiring offers more nuanced insights about the relationship between hiring and subsequent outcomes. This literature identifies a range of applicant information that appears to predict teacher performance (as measured by evaluations and student achievement gains) and retention. Effective predictors include screening rubrics (Goldhaber et al., 2017), centralized screening procedures (Bruno & Strunk, 2019; Jacob et al., 2018), machine-learning based measures of work histories (Sajjadiani et al., 2019), commercial screening tools that assess attitudes, beliefs, habits, and personality traits (Chi & Lenard, 2023), and commercial tools that assess cognitive ability and content knowledge (Rockoff et al., 2011). Structured ratings from professional references—the focus of this paper—have also shown promise as predictors of employee performance and retention (Goldhaber et al., 2024; Goldhaber & Grout, 2024).

Although the literature on teacher hiring suggests a range of applicant information shows promise, it rarely addresses a related, important question: if useful information on applicants exists, do hiring managers use it? Sometimes, the answer appears to be no. Jacob et al. (2018), for example, study a centralized applicant screening process in the District of Columbia Public Schools, which included written assessments of content knowledge, interviews, and scored teaching auditions. Conditional on being a recommended applicant, they find this screening information was weakly associated with the probability of being hired. Other studies suggest information on employees *can* inform managerial decisions, albeit after a hiring decision has been made. For example, Rockoff et al. (2012) conduct a randomized experiment in New York City Public Schools in which principals received measures of teacher performance based on student test scores. They found providing the information increased turnover for teachers with low performance and resulted in small gains in student test scores. To the best of our knowledge, there is no comparable study that focuses on the use of applicant information and hiring decisions. We address this gap in the literature using an experiment to assess whether hiring managers use low-cost information that offers signals of an applicant's potential as a future employee.

3. Study Setting and Information Experiment

Spokane Public Schools ("Spokane"), the setting for this study, is the largest school district in eastern Washington and the second largest in the state, serving approximately 29,000 students and employing roughly 2,000 teachers. Spokane uses an applicant tracking system (ATS) to post job openings, accept job applications, and manage the hiring process. Most job postings in Spokane are for specific positions (e.g., Teacher, Grade 2, at Lincoln Elementary), but some job postings are pooled (e.g., Teacher, K-3, Multiple Openings). Prior to applying to specific job openings, applicants to Spokane create a profile in the ATS. The profile includes information about an applicant's work history and credentials, and documents such as resumes and personal statements. Spokane also asks applicants to provide contact information for at least three professional references.

When an applicant applies for a job, their application is screened by central HR to verify that they meet the minimum qualifications for the position (i.e., having the necessary credentials and subject area endorsements). The applications of qualified applicants are advanced to school-level hiring managers, who use a standardized screening rubric to score applicants on a series of criteria. The highest-scoring applicants are offered an interview.[2] Because most job postings in Spokane are for specific positions, it is common for applicants to apply for multiple positions. In our analytic sample (described below), the median applicant applies to two positions and a quarter of applicants apply to 5 or more positions.

In prior work with Spokane, we studied the relationship between a screening rubric used by principals (and their school-based hiring teams) to identify which applicants to interview in person and teacher outcomes. We found that the screening scores were predictive of both value-added to student achievement in math and retention in the district (Goldhaber et al., 2017). Discussions with Spokane hiring managers revealed that the screening scores tended to be heavily informed by applicants' letters of recommendation. Although letters of recommendation were identified as an important source of information, hiring managers saw them as having important limitations. In particular, as noted by Goldhaber et al. (2024), "interpreting letters of

[2] The screening criteria are Education/Licensure, Related Work History, Cultural Competency, Reference Letters, Additional Information. Prior to 2019, the screening rubric also included the following criteria: Interpersonal Skills, Instructional Skills, Classroom Management, Flexibility, and Preferred Qualifications.

recommendation was challenging because these letters often failed to cover topics that mattered to hiring decisions (e.g., the applicant's classroom management skills), the fact that they were overwhelmingly positive required hiring officials to 'read between the lines' to get a nuanced assessment of applicants" (p. 2). We also found evidence that ratings hiring managers created based on the letters had low levels of reliability on some dimensions of prospective teacher skills (Martinkova et al., 2018). A natural question was whether better information could be collected directly from applicants' professional references.

3.1 Reference ratings

To address this question, we worked with Spokane to make a slight modification to their application process. As noted above, prior to applying to specific job openings, applicants to Spokane create a profile in the district's ATS and are asked to provide contact information for at least three professional references. The references are contacted by email via the ATS and asked to submit a confidential letter of recommendation using an online form.

Starting in June 2015, we worked with Spokane's ATS provider to redirect references to an online survey form following the submission of a letter of recommendation. We asked references to rate the applicant 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 are: *Challenges Students*, *Classroom Management*, *Working with Diverse Groups of Students*, *Interpersonal Skills*, *Student Engagement*, and *Instructional Skills*.[3] We used the following relative percentile ratings categories: *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*. The references were also asked to

[3] The survey form is shown in Appendix Figure A1 and the evaluation criteria are described in greater detail in Table A1 in the Appendix.

rate the applicant *Overall* using the same ratings categories, to indicate the competencies in which the applicant is *Strongest* and *Weakest* and were given space to provide open-ended comments. It was communicated to references that any information provided would remain confidential (i.e., would not be visible to the applicant).

We concentrated the ratings categories at the upper end of the distribution to give references the option to describe applicants positively without always selecting the top category.[4] This approach was successful in the sense that the ratings exhibit substantial variation: no more than 32% of ratings fall into any particular category. But the distribution of the ratings collected from references skews high. As shown in Figure 1, over half of the ratings on the *Overall* criterion indicate that applicants are *Among the best (top 1%)* or *Outstanding (top 5%)*. We observe very similar distributions for the six individual ratings criteria (see Figure A2 in the Appendix). This is unsurprising given findings from prior work which demonstrated that the ratings criteria were highly correlated with one another and loaded onto a single underlying factor (Goldhaber et al., 2021).

3.2 Information experiment

Our information experiment started in April 2018 when we began making the reference ratings information available to hiring managers for a randomized subset of applicants. When we received a rating of an applicant for the first time, we performed a virtual coin flip (with 50/50 odds) to determine the applicant's status as "rating-provided" or "rating-withheld." We randomized at the applicant level due to the structure of Spokane's application process; in particular, as described above, the applicant creates one profile that is used for all job

[4] It is common in evaluations of teachers and prospective teachers for evaluations to exhibit little variation (e.g., Goldhaber et al., 2022; Kraft and Gilmour, 2017).

applications.[5] Any additional information (including reference ratings) added to an applicant's profile is visible to all Spokane hiring officials, making it problematic to randomize at the job level. The experimental status of applicants (rating-provided vs. rating-withheld) was constant for the duration of the data collection period, April 2018 through August 2022.

The ratings information was provided to hiring officials as a one-page PDF report appended to the letter of recommendation submitted by the reference (see Figure 2), which is visible to hiring managers as part of the applicant's profile. The great majority of applicants in our analytic sample have multiple reference ratings (78%) and over half have three or more reference ratings. Each PDF report displays the categorical rating for each criterion, the competencies identified by references as the applicant's *Strongest* and *Weakest*, any additional comments provided the reference, and information about the reference including their name, email address, and relationship to the applicant (e.g., principal, colleague, or university supervisor). We also generated PDF reports for applicants for whom ratings were withheld with the following text inserted in place of the ratings information and comments: "For research purposes, professional reference ratings have been withheld for this applicant. Withholding status is determined at random and is not indicative of the quality of the applicant."

There was one modification to the process that occurred during the experiment. Ahead of the 2019 hiring year, Spokane switched to a new ATS provider whose system did not readily accommodate the solicitation of confidential letters of recommendation from applicants' references (instead, applicants were instructed to collect and upload letters to their profiles themselves). This required us to contact references directly via email with a request to complete

[5] Half of the applicants in our analytic sample (described below) applied to 2 or more positions.

the reference ratings survey and upload the PDF reports to applicants' profiles.[6] The primary difference in the provision of reference ratings information under the new ATS was that the PDF reports were no longer appended references' letters of recommendation (which were no longer confidential). Rather, the reference ratings (which remained confidential) were uploaded to applicant profiles as separate documents.[7] The collection and provision of reference ratings information concluded in September 2022.[8]

3.3 Job application and teacher outcome data

We are interested in understanding whether the information in the reference ratings influences hiring decisions. Accordingly, our analytic sample is anchored by job application records associated with classroom teaching positions during the period of the intervention (April 2018 through August 2022). Classroom teaching positions are identified based on job titles and exclude positions such as counselor, therapist, librarian, subject coach, summer school, and virtual learning. We also exclude job postings for which we do not observe at least one applicant with a school-level screening score and at least one applicant with a status of *hired*.[9] The ratio of unique applicants to openings for the jobs in our study sample (as indicated by the number of

[6] The shift to contacting applicants' references by email, rather than within the ATS, resulted in a lower response rate. About 6% of the emails sent to references were bounced back as undeliverable. Overall, we received a survey response for 69% of applicant-reference pairs and 92% of applicants had at least one response.
The PDF reports were assigned a document classification of "confidential", meaning that they were not visible to applicants.

[7] Unfortunately, we are not able to track whether and how often each reference rating document is accessed by hiring managers.

[8] During the course of the project, we experienced two significant disruptions to the collection of provision of reference ratings information. First, in March 2019, Spokane largely stopped posting job openings due to budget shortfalls brought about by a change in the state education funding landscape. Second, in March 2020, the Covid-19 pandemic disrupted all school operations, including hiring. Therefore, a relatively small number of reference ratings were collected during the 2019 and 2020 hiring years.

[9] District rules require that prior to the consideration of external candidates, school hiring managers must interview the two most senior properly certified employees requesting a transfer to a posted position. These senior transfer applicants are not visible to us in the data generated by the ATS and when hired, we do not observe any school-level screening scores or hiring outcomes for that job positing.

hires) is roughly 6: 1.[10] However, we also observe substantial variation in the ratio of applicants to openings by position type; for example, we see regular elementary-level teacher jobs drawing roughly 14 applicants per opening versus 5 applicants per special education opening.[11]

The job application records are linked to the reference ratings data using unique applicant IDs. We keep reference rating-job application links where the reference rating was generated no more than a week after the submission of a job application and no more than a year prior to the application date. These date restrictions are intended to ensure that reference ratings are reasonably current and are likely to have been uploaded to an applicant's profiles prior to the school-level screening stage of the hiring process.[12] As shown in Panel A of Table 1 (column 1), we observe 2,501 unique applicants who submitted 11,268 job applications to 462 job postings. The median applicant has three reference ratings (the mean is 2.8) and we observe 32,812 reference rating-job application links. Note that we treat individuals observed in different hiring years as distinct applicants because an applicant's qualifications will tend to change from one year to the next, particularly regarding work experience (14% of applicants in the analytic sample appear in multiple years).

Table 1 reports the number of observations at three key points in the hiring process. First, we observe the submission of an application (column 1). These applicants are subject to a district screening protocol that determines whether they have the necessary endorsements and credentials for the position in question. Second, we observe when an applicant is screened by a

[10] Note that the calculation of the 6:1 ratio includes applicants who are not associated with any reference ratings and therefore differs from the applicant to opening (as implied by the number hires) ratio suggested by Table 2, which is restricted to applicants associated with one or more reference ratings.
[11] In contrast to the 6:1 ratio, these calculations treat applicants who apply to multiple position types (e.g., elementary and SPED) as distinct resulting in an overall applicant to opening ratio of 8:1.
[12] Jobs are typically posted for at least five business days and applicants are screened by district HR prior to school-level screening. Therefore, it is highly likely that a reference rating collected within seven days of the submission of an application would be visible on an applicant's profile prior to the school-level screening stage of the hiring process.

school-level hiring manager (typically the principal) using a standardized rubric (column 2). As noted above, the standardized screening rubric is used to determine which applicants to interview in person.

We link school-level screening scores to job applications using applicant and job IDs. And third, we observe when an applicant is hired for position (column 3). We can see in Panel A of Table 1, that 70% of applicants advance to the school-level screening stage of the hiring process (1,759/2,501) and that 18% of applicants are hired (444/2,501). At the application level, however, we observe far lower rates of advancement in the hiring process with just 38% advancing to school-level screening and 4% being hired. In several cases, we observed candidates with a status of *hired* for multiple positions, resulting in a slightly larger number of hires at the application level than the applicant level (448 vs. 398).

We link hired applicants to personnel data maintained by Spokane.[13] These data allow us to generate measures for two post-hire outcomes: performance evaluations and retention in the hiring school.[14] We focus on the outcomes of hired applicants because the school-level screening scores are job specific and as described below, our analysis will involve modeling the relationship between principals' ratings of applicants and teacher performance and retention. We

[13] Unfortunately, applicant information, like applicant-reported experience and degree level, was only available in the year in which Spokane utilized WinOcular, so the sample sizes are small (525 applicants). We observe relatively small changes in experience as individuals progress through the hiring pipeline and into the two in-service samples. But surprisingly, hired teachers have about a year less of reported experience (5 years) than the overall applicants pool (6 years) and are likely less likely to hold an advanced degree.

[14] In prior work, we also examined the relationship between reference ratings and teacher performance as measured by value-added use reference ratings collected during 2015 to 2018 (Goldhaber et al., 2024). Unfortunately, the pandemic resulted in an extended disruption to standardized testing in Washington State so estimates of teacher value added are unavailable for most years in the current study, resulting in small sample sizes. While we omit value added estimates from the current study due to the resulting lack of precision, we found moderate correlations between performance evaluation ratings and teacher value-added of 0.23 (math) and 0.21 (reading) in our prior work.

were able to link 398 and 428 of the 444 hired candidates to performance evaluation and retention outcomes, respectively.[15]

In Panel B of Table 1, we present the distribution of references' ratings on the *Overall* criterion at the reference rating-job application observation level. As expected, since professional references are both writing letters of recommendation and responding to the reference ratings survey, we find that the reference ratings are positively associated with hiring outcomes. For example, reference ratings of *Outstanding* and *Among the Best* are slightly over-represented among hired applicants (by between 1.5 and 3.4 percentage points) whereas ratings of *Very Good* and *Average* are slightly under-represented among hired applicants (by between 4.2 and 6.7 percentage points). There is little difference, however, in the distribution of ratings of all applicants (column 1) and those advancing to the school-level screening stage of the hiring process (column 2). This is consistent with the district-level screening process being focused on whether the applicant meets the minimum qualifications for the position (i.e., they have the appropriate credentials and subject-area endorsements) and clears a background check rather than differentiating applicant quality.

As noted above, the reference ratings data were provided to applicants on a randomized basis, with provided/withheld status determined at the applicant level. As such, we would expect applicant characteristics and reference ratings to exhibit similar distributions in the rating-withheld and rating-provided sub-samples. As shown in Table 2, we observe a similar distributions of reported experience in the two sub-samples as well as similar proportions of

[15] We rely on Spokane's personnel data to determine performance evaluation and retention outcomes. There are several reasons that we were not able to link some hired candidates to outcomes: the hired candidate may have left prior to the end of the school year and was not evaluated; the candidate was a tenured internal hire and was not subject to a comprehensive performance evaluation; we were unable to link the candidate to the personnel data by matching on name; job offer was rescinded or declined.

internal applicants. However, the proportions of candidates rated *Below Average* or *Average* are slightly higher among the rating-withheld candidates than among rating-provided candidates (by 0.005 and 0.016, respectively), and the difference is statistically significant. We observe qualitatively similar patterns for the individual ratings criteria (see Table A2 and Figure A2 in the Appendix), though there is also interesting variation across the individual criteria. For instance, virtually no references respond that they have “no basis for judgment” on the “Interpersonal Skills” criterion, whereas over 6% of respondents say this about “Classroom Management”.

3.4 Measures

Here we describe three measures used in the empirical analyses outlined in Section 4 below: a summative reference ratings measure, standardized school-level screening scores, and performance evaluations. Each of these measures applies a graded response model to the data. As described in Goldhaber et al. (2024), the graded response model which is suited to ordered categorical data and allows criteria to vary in difficulty and discrimination. We estimate the probability of observing a rating level k or higher on criterion c of the rating of applicant or teacher i in year t (where the rating is the reference rating, school-level screening score or performance evaluation score):

$$\Pr(Rating_{itc} \geq k|\theta_{it}) = \frac{\exp\{a_c(\theta_{it}-b_{ck})\}}{1+\exp\{a_c(\theta_{it}-b_{ck})\}}, \qquad (1)$$

where a_c represents the discrimination of criterion c, b_{ck} is the kth cut point of criterion c, and θ_{it} is the latent quality represented by the GRM estimates. We use estimates of θ_{it} as summative measures of an applicant/teacher’s quality/performance, as described below.

Reference ratings

As described above, we collected categorical ratings of applicants from their references on series of criteria. In some components of our empirical analysis, we will represent these ratings as categorical variables, consistent with how the reference ratings information is presented to hiring managers in applicants' profiles. For ease of interpretation, and in the interest of generating a summative reference ratings measure that incorporates information from ratings on each of the six criteria (which are strongly correlated with one another), we apply the graded response model described by equation 1 to these categorial data. Standard errors are clustered at the applicant-year level (remember that most applicants have ratings from multiple references) and ratings of *No basis for judgment* are treated as missing values because they do not fit in the context of ordered categorical data.[16] While ratings of *No basis for judgment* are excluded from the likelihood estimation, we do not exclude entire observations. Rather, we obtain estimates of θ_{it} for each observation in the sample using the criteria that *are* rated and standardize the summative reference ratings measure $\sim(0,1)$ by hiring year.

School-level screening scores

As noted above, school-level hiring managers use a standardized screening rubric to score applicants on a series of criteria. Prior to Spokane's transition to a new ATS in 2019, this rubric consisted of 10 criteria scored on a scale of 1 to 6. These criteria were *Certificate and Education*, *Experience, Training*, *Cultural Competency*, *Preferred Qualifications*, *Letters of Recommendation*, *Interpersonal Skills*, *Instructional Skills*, *Flexibility*, and *Classroom Management*. The screening forms associated with a particular job posting were emailed or faxed

[16] As shown in Table A2, the frequency of ratings of *No basis for judgment* range between roughly 0.5% (*Interpersonal Skills*) and 6.1% (*Classroom Management*).

to HR once the school-level screening process for that job was completed and we generated digital records of each screening form, including applicant names and job IDs.

To generate a summative screening score for the 2018 data, we apply the graded response model described in equation 1 to obtain estimates of θ_{it}. As with the reference ratings, we exclude missing values from the likelihood estimation and standardize the summative ratings measures $\sim(0,1)$. There are two sources of missing values. In some cases, the rater opted not to evaluate candidates on one or several criteria (about 10% of screens). And more commonly, when completed screening forms were sent to HR, the sender failed to make two-sided scans of the documents, resulting in missing values for scores on the second page of the rubric (about 35% of screens). Both types of missing values tend to be consistent within-job: if the score on a criterion is missing for one applicant to a particular job it tends to be missing for every applicant to that job.

When Spokane transitioned to their current ATS in 2019, the school-level screening process was fully incorporated into the ATS's online platform. This change was accompanied by a modification to the screening rubric, which asks hiring managers to rate applicants on the following criteria on a scale of 1 to 5 at increments of 0.5: *Education/Licensure*, *Related Work History*, *Cultural Competency*, *Reference Letters*, and *Additional Information*. As with the 2018 screening score data, we apply the graded response model in equation 1 to obtain estimates of θ_{it}.

Performance evaluations ratings

Spokane teachers are evaluated under Washington State's Teacher and Principal Evaluation Program (TPEP) using a state-approved rubric that rates them as belonging in one of four performance categories on eight different competencies.[17] We estimate the graded response model described in equation 1 to obtain estimates of θ_{it}, which serve as a summative performance evaluation measure.

4. Empirical Approach

We are interested in whether structured ratings of teacher applicants collected from their professional references have the potential to improve hiring decisions. Specifically, we assess whether reference ratings influenced hiring decisions when provided to hiring managers and examine their potential improving hiring outcomes.

4.1 The use of professional reference ratings

To analyze whether reference ratings are used by hiring managers, we model the relationship between reference ratings and applicants' progression through the hiring process, leveraging the fact that the ratings were made available to hiring managers for a random subset of applicants (as described in Section 3.2).

First, we examine whether reference ratings influence applicants' progression to the school-level screening stage of the hiring process, indicating that they have cleared Spokane's central screening protocol. Specifically, we estimate the following model:

$$f\left(SchoolLevel_{ij}\right) = \alpha_0 + \alpha_1 \overline{RR}_{ij} + \alpha_2 P_i + \alpha_3 \left(\overline{RR}_{ij} * P_i\right) + \alpha_4 I_i + \gamma_j + \varepsilon_{ij}, \qquad (3)$$

where $SchoolLevel_{ij}$ is an indicator equal to one if applicant i advances to the school-level-screening stage of the hiring process, $\overline{RR}_{ij}$ is the average the reference ratings associated with

[17] For more information about TPEP, see https://ospi.k12.wa.us/educator-support/teacherprincipal-evaluation-program (accessed June 5, 2024).

applicant i for job j, P_i is an indicator that the reference ratings of applicant i are provided, I_i is an indicator that applicant i is an internal applicant, and γ_j is a job fixed effect.[18]

The model is estimated as either a logit model or linear probability model, with standard errors clustered at the applicant level. In our preferred specification, we estimate the model at the *applicant-job level* with a job fixed effect such that the estimated coefficients are identified by within-job variation. An advantage of this specification is that it aligns well with the actual hiring process, in which the qualifications for a position and accordingly, assessments of applicant quality, are job specific. This is also consistent with the notion that hiring managers carry out some type of when evaluating information multiple references. In addition to our preferred specification, we estimate the model at the *applicant level* using the average summative reference rating $\overline{RR}_i$, which is consistent with the level of randomization. In this specification, the dependent variable is an indicator that the applicant advanced to the school-level screening stage of the hiring process for *any* job. We cannot include job fixed effects in the *applicant-level* specification but add controls for the amount of competition for those jobs, the number of jobs the applicant applied to, a year fixed effect, and the number of reference ratings associated with the applicant. To assess robustness, we also estimate the models at the *applicant-job-reference rating level* (using RR_{ijr}). An advantage of estimating the model at this level is that it allows us to represent the professional ratings in the form in which the hiring managers view them (rather than, for instance, averaging across raters for applicants who have multiple structured ratings). Specifically, we can represent the reference ratings—on the *Overall* criterion and on each individual criterion—as categorical variables.

[18] $\overline{RR}_{ij}$ is the average of the summative measure derived from the estimation of the graded response model described in Section 3.4.

Next, we model whether reference ratings influence hiring managers' assessments of applicants. We leverage the fact that the school-level screening scores are representative of hiring managers' assessments of applicants given the information available in the candidates' application profiles and the fact that the reference ratings are withheld for a random subset of applicants. If the reference ratings are influencing hiring managers' assessments of applicants, we would expect the relationship between the reference ratings and the school-level screening scores to be stronger for *provided* ratings than for *withheld* ratings. To test this proposition, we estimate the following linear regression model:

$$Screen_{ij} = \alpha_0 + \alpha_1 \overline{RR}_{ij} + \alpha_2 P_i + \alpha_3 \left(\overline{RR}_{ij} * P_i\right) + \alpha_4 I_i + \gamma_j + \varepsilon_{ij} \ , \qquad (4)$$

where $Screen_{ij}$ is the school-level screening score for applicant i to job j and the other variables are as defined above. We estimate alternative specifications consistent with those described for model (3) above. Finding that $\hat{\alpha}_3$ is statistically significant would provide causal evidence that the reference ratings are influencing school-level hiring managers' assessments of applicants.

Finally, we predict the probability that an applicant is hired. We modify equation 3, replacing the dependent variable with $Hired_{ij}$, an indicator that applicant i was hired for job j. We estimate the model conditional on applicant i having advanced to the school-level screening stage of the hiring process for job j, meaning that they meet the minimum qualifications for the position. And again, we estimate alternative specifications consistent with those described for model (3) above. We do not observe job offers in Spokane's current ATS, but earlier work showed that nearly all applicants (about 95%) who receive a job offer accept it (Goldhaber et al., 2017), and declined offers remain uncommon according to Spokane HR personnel (personal communication, June 26, 2024).

4.2 Professional reference ratings and information

Prior work has demonstrated that reference ratings are predictive of teacher performance and retention outcomes (Goldhaber et al., 2024; Goldhaber and Grout, 2024), but there is limited evidence about the degree to which the informational content of the reference ratings may be used to substitute or supplement the information that is already collected about teacher applicants. To examine the degree of informational overlap between reference ratings and existing applicant information we use reference ratings and school-level screening scores to predict teacher outcomes. We leverage the fact that school-level screening scores are representative of hiring managers' assessments of applicants given the information available in the candidates' application profiles. If the reference ratings provide additional information, they should be predictive of teacher outcomes even when controlling for school-level screening scores. Among applicants with *provided* ratings, it is possible that the information in the reference ratings is incorporated into the school-level screening scores. Therefore, we focus on the applicants whose reference ratings are *withheld*.

In modeling the relationship between reference ratings, screening scores, and teacher outcomes, we worry about sample selection bias. The screening scores are used to determine advancement in the hiring process and the withheld reference ratings are likely to be correlated with factors predictive of being hired. To obtain unbiased estimates, we estimate a Heckman selection model. Following Goldhaber et al., (2024) we calculate two instrumental variables that measure the amount of competition faced by applicants during the hiring process.

1) *Quantity of competition* – calculated at the job level as $(number\ of\ competitors)/(number\ of\ openings)$

2) *Quality of competition* – calculated at the job level as the 75th percentile of the mean summative reference ratings of the competing applicants

The assumption is that the amount of competition applicants face is predictive of their being hired but is exogenous to the performance and retention outcomes of hired applicants.[19]

We specify the first stage of the Heckman model as follows:

$$Hired_{ij}^{*} = \beta_0 + \beta_1 Screen_{ij} + \beta_2 \overline{RR}_{ij} + \beta_3 I_i + \beta_4 S_j + \gamma_t + \beta_5 Z_{ij} + \varepsilon_{ij} \qquad (4)$$

where $Hired_{ij}$ is an indicator that applicant i is hired for job j, $Screen_{ij}$ is the school-level screening score, and $\overline{RR}_{ij}$ is the average summative reference rating as described above. I_i is an indicator that applicant i is an internal applicant, S_j is a vector of controls for the subject area of the job indicating whether the position is for grade teacher, English language arts, STEM, special education or other. γ_t and δ_s are school and year fixed effects. The instruments—quantity of competition and quality of competition—are in the vector Z_{ij}. Then, letting

$$Hired_{ij} = I\left(Hired_{ij}^{*} \geq 0\right), \qquad (5)$$

we estimate the conditional model:

$$f\left(TPEP_{ij} \middle| Hired_1\right) = \beta_0 + \beta_1 Screen_{ij} + \beta_2 \overline{RR}_{ij} + \beta_3 S_j + \gamma_t + \varepsilon_{ij}, \qquad (6)$$

where $TPEP_{ij}$ is the summative performance evaluation measure (derived in Section 3.4 above) for the hired applicant i for job j.

We estimate an equivalent model for school-level retention outcomes, where the dependent variable is an indicator equal to one if the applicant hired in year t is retained at the hiring school in year $t + 1$. As in the TPEP model, we include a subject area control to account

[19] We estimate standard Heckman models for the TPEP and retention outcomes using Stata's *heckman* command. We obtain qualitatively similar results when estimating a Heckit model for retention outcomes using Stata's *heckprobit* command.

for the fact the teachers in different subject areas exhibit different rates of retention (e.g., Nguyen et al., 2020). The model is estimated as either a logit model or linear probability model with standard errors clustered at the applicant level.

5. Results

We describe the extent to which structured ratings of teacher applicants ("reference ratings") from their professional references might improve hiring decisions below. We begin by examining whether the provision of reference ratings influences hiring managers' assessments of applicants and hiring decisions. To add context to our findings, we then examine the degree to which the informational content of the reference ratings overlaps with existing applicant information. In our preferred specifications, we predict hiring and teacher outcomes using applicants' average summative reference ratings at the applicant-job level, but also consider models estimated at the applicant and applicant-job-rating levels. For ease of interpretation, the summative reference ratings, screening scores and TPEP evaluation measures are standardized $\sim(0,1)$ and binary outcome models are estimated as linear probability models.

5.1 Findings: the use of professional reference ratings

To examine whether reference ratings influence hiring managers' assessments of applicants and hiring decisions, we model the relationship between the provision of reference ratings to hiring managers for a random subset of applicants and the progression of applicants through the hiring process. Specifically, we analyze whether reference ratings affect the following: clearing Spokane's central screening protocol (so that applicants are advanced to school-level screening); school-level screening scores; and the likelihood of being hired. In our preferred specification, we estimate the model with job fixed effects so that the coefficients are identified by within-job variation in reference ratings.

To test whether reference ratings information is influencing the hiring process, we include the reference ratings in the models along with an indicator that the reference ratings were $provided$ and the interaction term $Provided * Reference\ Ratings$. Note that it is natural for the reference ratings to be correlated with applicants' progression through the hiring process whether or not they influence the hiring process since the raters are also writing letters of recommendation. A significant coefficient on the *provided* indicator would suggest that the provision of reference ratings influences the hiring process, independent of the level of the reference ratings. This could occur if the provision of ratings tended to reduce the level of perceived uncertainty around applicants or systematically cast applicants in a positive (or negative) light. A significant coefficient on the interaction between provided and rating would indicate that providing reference ratings to hiring managers influences their assessments of applicants and/or hiring decisions.

We present findings from models estimated at the *applicant-job level* (in panel A) and at the *applicant level* (in panel B) in Table 3.[20] In panel A, we present results with and without job fixed effects since we do not know how hiring officials consider the ratings, i.e., they might compare applicants across (similar) jobs or only compare applicants for the particular job that is open. The *applicant-level* results in panel B are presented with and without controls accounting for the level of competition.[21] At the *applicant-job level*, reference ratings are generally predictive of hiring process outcomes but are not predictive of advancing to school-level screening when the model includes job fixed effects. A one standard-deviation change in average

[20] Results estimated at the *applicant-job-rating level* (presented in Table A4 in the Appendix) are qualitatively similar to the *applicant-job level* results presented in pane A of Table 3.

[21] The controls for the amount of competition include the instrumental variables measuring the quantity and average quality of competition described in Section 4.2 (averaged across jobs), the number of jobs the applicant applied to, the number of ratings associated with the applicant, and hiring year indicators.

reference rating is associated with a 22% of a standard deviation increase in school-level screening score and a 3.7 percentage point increase in the probability of being hired.[22] These findings are to be expected irrespective of whether reference ratings are being used because references are also writing letters of recommendation that are provided to hiring managers regardless of treatment status.[23]

Assignment to the treatment of having one's reference ratings provided to hiring managers is not generally predictive of hiring process outcomes. However, the coefficient on the indicator is positive for the school-level screening score and hiring models (in columns 3 to 6) and is marginally predictive of being hired when we do not include job fixed effects (column 5 of Panel A). The point estimates here are small; applicants whose ratings are provided are 1.5 percentage points more likely to be hired. As noted above, one possibility is that applicants whose reference ratings are provided are viewed more positively than applicants whose reference ratings are withheld because the ratings are generally so positive. For example, a rating of *Excellent (top 10%)* is a below-average rating given the distribution of ratings we observe but may cause an applicant to be viewed more positively when provided to hiring managers. Another possibility is that hiring managers perceive less uncertainty around applicants with provided reference ratings.[24]

[22] About 40% of applications advance to the school-level screening stage so this figure represents an increase in the likelihood of advancing of about 5 to 10 percent. The results from the hiring models (in columns 5 and 6) are estimated conditional on having reached the school-level screening stage of the hiring process but are very similar when all applicants are included in the regression model.

[23] Each of the models in Panel A includes a control for whether the applicant is an internal candidate. We find that internal candidates receive school-level screening scores that are roughly 26% of a standard deviation higher but are not significantly more likely to be hired. Being an internal candidate is also predictive of school-level screening scores in the applicant-level models in Panel B, and of being hired.

[24] As described in Section 3.2, the provision of reference ratings is randomized at the applicant level. Therefore, for any given job, the hiring manager will be provided reference ratings information for some subset of applicants. We find that when fewer than half of the applicants for a job are in the treatment group (applicants with provided reference ratings), being in the treatment group is predictive of being hired (+2.6 percentage points), and no relationship when more than half of applicants are in the treatment group.

Aside from the above finding on the provision of ratings, we find little evidence in the *applicant-job level* models that the information contained in the reference ratings positively influences hiring managers' decisions. The coefficient on the interaction term $Provided * Reference\ Rating$ is close to zero and statistically insignificant in each *applicant-job level* model (Panel A), indicating the relationship between reference ratings and hiring outcomes is no stronger among applicants for whom the ratings are provided than among applicants for whom the ratings are withheld. The null results for the interaction term are somewhat imprecise; we can only rule out with 95% confidence effect sizes larger than 10.3 percent of a standard deviation change in school-level screening scores and a 1.7 percentage point increase in the probability of hire.[25]

We find a similar pattern of results when the models are estimated at the applicant level (in Panel B).[26] The only area of significant difference arises in the models predicting hiring outcomes (columns 5 and 6). When estimated at the applicant level, providing reference ratings significantly reduces the relationship between the reference ratings and the likelihood of being hired. As noted above, a possible explanation for this counterintuitive result is that weaker applicants whose reference ratings are provided may be viewed more positively than applicants whose reference ratings are withheld if hiring managers treat below-average ratings (such as *Excellent (top 10%)*) relatively favorably, particularly if a hiring manager had been exposed to relatively few reference ratings. We test this by allowing the relationship between the interaction

[25] One concern with the *applicant-job level* models is that outcomes may be correlated if an applicant applies to multiple positions at the same school: if the first application resulted in a low school-level screening score and no hire, this outcome is likely to reoccur in subsequent applications. As a robustness check, we estimate models restricted to the first observed pairing of an applicant and school-level screener and obtain very similar results.

[26] While not of primary interest, these models include controls for the number of raters and the jobs to which candidates apply. Both of these are statistically significant and positive. The inclusion of these variables, however, has little effect on the estimates on the reference rating or the interaction between provision of ratings and the reference rating.

term $Provided * Reference\ Rating$ to vary according to how many reference ratings have been made available to the school-level hiring manager for the position in question and find mixed support for the hypothesis.[27] Specifically, we generate an indicator that a hiring manager has been provided 5 or fewer reference ratings associated with the jobs in our analytic sample and find a negative and significant relationship between the interaction term $(Exposure \leq 5) * Provided * Reference\ Rating$ and the probability of hire. Conversely, we find a positive and statistically insignificant relationship on the interaction term $(Exposure > 5) * Provided * Reference\ Rating$.[28] However, this pattern does not hold up when the exposure threshold is increased to 10. Additionally, under this "viewed more positively" explanation we would expect to find a similar relationship between the provided reference ratings and school-level screening scores and as shown in Table 3, this is not the case.

A second possible explanation for the counterintuitive finding that providing reference ratings significantly reduces the relationship between the reference ratings and the likelihood of being hired is that it reflects sample selection. Specifically, stronger applicants may be selected earlier in the hiring cycle, and weaker applicants will tend to apply for more jobs and have a better chance of being hired later in the hiring cycle.[29] The dependent variable in the applicant-level model does not distinguish applicants hired into the first job to which they applied from applicants that find positions after multiple applications. When we restrict the applicant-level

[27] We can track this information for the 2019 to 2022 period, after Spokane switched to a new ATS provider and fully digitized its school-level screening score process.

[28] The model is estimated at the *applicant-job level* with controls for whether the applicant is an internal candidate, the quantity and quality of competition, and year indicators. The coefficient on $(Exposure \leq 5) * Provided * Reference\ Rating$ is -0.037^* and the coefficient on $(Exposure > 5) * Provided * Reference\ Rating$ is 0.001.

[29] Among applicants with at least one school-level screening score, the correlation between the number of jobs applied to an average reference rating is -0.173.

model to each applicant's first job application, we no longer find a significant relationship between the interaction term $Provided * Reference\ Rating$ and the probability of hire.[30]

The analyses discussed above use the summative measure derived from the graded response model. But this is not what the school-based hiring teams see when the ratings are made available. As shown in Figure 2, the categorical ratings are presented for each of the six evaluation criteria as well as the *Overall* criterion. When we estimate the models presented in Table 4, entering the reference ratings as categorical variables, we find very similar results.[31] The categorical ratings are predictive of school-level screening scores and hiring outcomes, but the provided reference ratings are not *any more predictive* of these outcomes than the withheld ratings. Predictive margins from the models for the *Overall* criterion are presented in Figure 3 (school-level screening scores) and Figure 4 (hiring outcomes), and equivalent plots for the specific rating criteria are presented in Figures A3 and A4 in the appendix.[32]

5.2 Findings: professional reference ratings and information

To analyze the informational overlap between reference ratings and existing applicant information, we predict teacher performance evaluations ("TPEP") and school-level retention using the reference ratings and school-level screening scores (for the sample of teacher applicants in which the reference ratings are withheld). As discussed in Section 4.2, we focus on the sample of teachers whose ratings are *withheld* and estimate a Heckman selection model to account for potential bias introduced by selection into the sample. The results from the Heckman

[30] The coefficient on the interaction term is -0.13 without job fixed effects and 0.022 with job fixed effects and the standard errors are similar to those of the interaction terms in columns 5 and 6 of Table 3 (Panel B).

[31] The models must be estimated at the applicant-job-rating level because the reference ratings are categorical variables and not amenable to averaging. We estimate separate regression models for each criterion because they are strongly correlated with one another and as documented in prior work, a factor analysis of the reference ratings showed that they load onto a single factor.

[32] We do not do this at the other levels of the analyses (applicant and applicant job levels) because the applicant ratings are categorical.

model are presented in Table 4, with the TPEP model in Panel A and the school retention model in Panel B; for comparison purposes, we show uncorrected for sample selection (OLS estimates) in column 3 of both panels. For ease of interpretation, we estimate the retention model as Heckman model rather than a Heckit model. We find qualitatively similar results under the Heckit model specification, which are available in Table A5 in the Appendix.

The average marginal effects for the first-stage selection models are presented in column 1 (recall, as discussed in Section 3, that the two outcome samples do not perfectly overlap). The instruments perform well in both models, with greater levels of competition corresponding to a significantly lower probability of selection. A 1 standard deviation change in the quantity and quality of competition measures correspond to roughly 3-percentage point and 1-percentage point decreases in the probability of hire. Because most applicants apply to multiple jobs, the job-level probability of being hired is quite low (about 4%).

We find the school-level screening scores are significantly predictive of selection into the sample: a one-standard deviation change in an applicant's screening score is associated with about an 8-percentage point increase in probability of being hired. The coefficient on reference ratings is small and statistically insignificant in the TPEP evaluation sample, but marginally significant in the school retention sample This is not an unexpected finding as even for the sample of teacher applicants where hiring officials do not see the ratings as the raters are also writing letters of recommendation which are seen by hiring officials.

The selection-corrected teacher outcome models are presented in column 2. The school-level screen scores have large (but imprecisely estimated) effects on classroom observation-based measures of performance (TPEP). The reference ratings are strongly and significantly predictive of TPEP scores, with a one-standard deviation change in reference rating associated with 41% of a standard deviation change in TPEP scores. We do not, however, find the reference ratings to be predictive of retention. While the coefficient on the Inversion Mills Ratio is statistically insignificant, we do see evidence of attenuation for school-level screening scores. The estimated coefficients in the selection corrected models are far larger than those obtained from the OLS models: 0.289 vs. 0.170 for TPEP outcomes and 0.114 vs. 0.091 for the school retention models. This finding is unsurprising given that the school-level screening scores are used to determine which applicants to interview in person.

The above findings suggest that the reference ratings provide information that is additional to that otherwise available in applicant profiles, i.e., they are significant in models that include the school-level screening score as a covariate.[33] In the next section we explore the extent to which the use of reference ratings might influence the composition of hired teachers.

6. Discussion of Scope for Change

Our findings suggest that reference ratings significantly supplement the information available in applicants' profiles (in terms of predicting teacher performance), but also that the information is not being used in expected or productive ways. In particular, as shown in Table 3, although there is evidence that the provision of reference ratings information is marginally predictive of being hired, we fail to find significant evidence that the ratings are being used to differentiate applicant quality. In this section we explore the extent to which using the reference ratings in more productive ways might improve the quality of the teacher workforce.

To explore the scope for change and improvement, we begin by considering how much overlap there is between the reference ratings and school-level screening scores. The greater the overlap between the two measures, the smaller the scope for change. We do this as follows: For

[33] Neither measure is statistically significant in the retention models. These results are available upon request.

each job, we rank the applicants under three different scenarios: 1) using only applicants' school level screening scores for each job; 2) using only the average of applicants' summative reference ratings measures associated with each job; and 3) using a weighted average of the school-level screening scores and average reference ratings, with the weights determined by the coefficients from the second stage of the Heckman model presented in Table 4. This weighted average approach maximizes predicted performance evaluations, holding other factors constant. When there is a tie for the top rank, we classify both applicants being the top-ranked candidate. 34 We characterize the two ranking scenarios as being in agreement when the top-ranked candidate is the same individual according to both ranking criteria.

In Table 5, we present the level of agreement between the different ranking scenarios of applicants and how frequently the top-ranked applicant under each scenario was the applicant hired by Spokane. For simplicity, we focus on the 302 jobs where only one candidate is hired (there are a total of 323 jobs and 378 hires with screening scores and TPEP outcomes in our analytic sample). We find that the top-ranked applicant according to the school-level screening score is frequently different than the top-ranked applicant according to the reference ratings. The two measures agree only 38% of the time. When ranking applicants according to the weighted measure, which incorporates both school-level screening score and reference ratings information, the rate of agreement with the school-level screening score is 63%. In other words, the weighted measure is identifying a different individual as the top-ranked applicant 37% percent of the time.

We also consider how often the top-ranked applicant is hired. Given the explicit use of the screening scores in the district's hiring process, screening scores are unsurprisingly the most predictive of eventual hires. But interestingly, the applicant with the top screening score is only

[34] There are 116 ties when using the screening score as the ranking criterion, six ties when using the reference ratings, and 1 tie using the weighted measure.

hired 55% of the time.35 This suggests the importance of the interview process, which typically involves 3 to 5 candidates who reach that stage because of their screening scores.36 It is only 42% of the time that a top-ranked applicant is hired when the weighted measure is used to determine applicant ranking, and only 30% when reference ratings alone are used. These findings provide evidence that there is scope for improvement in hiring, as the rankings vary a good deal according to the ranking criteria.

We next compare the predicted performance of actual hires to the predicted teacher performance of teachers hired under alternative ranking scenarios. We calculate predicted TPEP estimates using the estimates derived from of the Heckman model presented in columns 1 and 2 of Table 4. We consider each job for which the hired applicant has a reference rating, screening score, and observed TPEP outcome (unlike the ranking exercise above, we do not exclude jobs with multiple hires).37 As above, we generate three sets of rankings for the applicants associated with each job: according to the school-level screening scores, reference ratings, and a weighted average of the two measures.

We build three alternative slates of hires using the different ranking criterion as follows. Starting with the school-level screening score criterion, we begin by sorting the jobs according to the first application date associated with each job. Beginning with the earliest job date, we mark an applicant as hired into the job if they are the top-ranked candidate according to the school-level screening score criterion and remove that applicant from the applicant pools of any other jobs to which they had applied at a future date. When multiple candidates tied for the top rank,

[35] The figure of 55% is not driven by ties in the screening score rankings. As we note above, when two applicants tie for the top rank, both are classified as "top-ranked" such that either one would count as a top-rank hire.

[36] Unfortunately, we do not directly observe which candidates are interviewed.

[37] There are 323 jobs where we the hired applicant(s) has one or more reference ratings, a school-level screening score, and TPEP outcome. Associated with those 323 jobs are 378 hired applicants and 1,528 unique applicants and 3,397 applications that advanced to the school-level screening stage of the hiring process.

the hired applicant is selected randomly. We repeat this process cycling through all 323 jobs. We then calculate the average predicted performance (TPEP) of the hired applicants. We repeat this same process using the reference ratings criterion and the weighted average criterion.

The results from this exercise are presented in Table 6. When selecting on the reference ratings rather than the school-level screening score, there is almost no change in the predicted performance of future teachers, suggesting a fair degree of substitutability between the two measures. And while we find that average predicted performance when selecting on school-level screening scores or reference ratings is about 20% of a standard deviation higher than the average predicted performance of the actual hires, we do not infer that this is the scope for improved hiring. The reason is that hiring officials have opportunities to learn about applicants during the interview stage of the process and that information (which we do not observe) may also be predictive of performance.

Using both the school-level screening score and the reference rating together in the weighted-average is predicted to yield teacher performance that is about 10% of a standard deviation higher than when using the screening scores or reference ratings alone. This is likely an upper bound on the scope for improvement, since there are some applicants under each ranking criterion that might make it to the interview process and not be hired based on what is learned at that stage of the process. Still, the effect size is substantial—about half the difference in the average performance evaluation scores of teachers with one year of experience versus teachers with two years of experience. This effect size is similar to the returns to early career experience on observational evaluations found in other educational contexts (Bartanen et al., 2024).

7. Conclusions

Efficiently matching applicants to job openings is a well conceptualized problem in personnel economics. But as noted by Hoffman and Stanton (2024), “So far, economists have performed relatively little work in understanding the value of adopting different hiring procedures” (p. 46). Our experimental findings contribute to the literature on personnel economics by providing empirical evidence on whether a specific hiring practice—providing ratings of applicants collected from their professional references—influences hiring decisions.

Despite evidence that information from professional references could improve hiring decisions, we find that hiring managers fail to use it. This finding is broadly consistent with nonexperimental evidence on the use of teacher applicant information (Jacob et al., 2018).

Not using the reference information appears to be a missed opportunity for hiring managers and schools. We find that reference ratings can efficiently capture novel information about a prospective teacher’s future potential. As noted in Goldhaber et al. (2024), we implemented an automated system for collecting ratings of applicants from references at a one-time cost of $2,000 and without much burden on administrators. It is reasonable to think that most ATS providers could incorporate the collection and provision of reference ratings information into their platforms at little expense.

The disconnect between the potential for reference ratings to better inform hiring decisions and the lack of evidence that hiring managers use them raises questions about how to more effectively encourage their use. One potential area for improvement is to modify how the reference rating information is presented to hiring managers. We provided the reference ratings information “as is.” If an applicant is rated as “Excellent (top 10%)”, that is what is shown to hiring managers. But as we show (see Figure 1), ratings tilt toward the top of the scale, which may be misleading to hiring managers. For instance, a rating of “Excellent (top 10%)” is actually a below-average rating. Standardizing the ratings prior to presenting them to hiring managers would provide them with more context, potentially increasing the ratings practical usability. Another way to provide more context to hiring managers would be to leverage the fact that many references (14%) have rated multiple applicants. In these cases, the ratings of an applicant could be juxtaposed against the ratings of other applicants generated by the same reference.

Discussing the generalizability of our findings is somewhat of a speculative exercise. On the one hand, the potential utility of reference ratings may be overstated if most school districts have limited applicants, unlike Spokane. While the ratio of teacher applicants per open job (roughly 6:1) is similar to what is reported in other studies on teacher hiring practices (e.g., Jacob et al., 2018; James et al., 2024), there are reports of widespread teacher shortages elsewhere (Nguyen et al., 2022). Districts only benefit from additional information about applicants if there are multiple applicants to choose from. Unfortunately, there is limited information on how qualified applicants are distributed across the teacher labor market (Bleiburg and Kraft, 2022; Nguyen et al., 2022).

On the other hand, studying the use of reference ratings in a district like Spokane may understate their potential utility because Spokane had a hiring process that was already relatively sophisticated. For example, as described above, Spokane uses a standardized screening rubric at the school level to select which applicants to interview in person, and it has been shown to be predictive of teacher effectiveness and retention (Goldhaber et al., 2017). The potential for reference ratings to improve hiring decisions may be greater among districts that are less systematic about collecting information about teacher applicants. However, that assessment presumes that reference ratings could be used productively—something that was not demonstrated by the provision of reference ratings information in Spokane. Our findings squarely support the notion that better applicant screening could be an important tool for improving teacher hiring, but there is still much to be learned about the ways in which school systems can improve teacher quality through better applicant screening and selection.

References

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. (2015). Do First Impressions Matter? Predicting Early Career Teacher Effectiveness. *AERA Open*, 1(4), 1–23. https://doi.org/10.1177/2332858415607834

Backes, B., Cowan, J., Goldhaber, D., Theobald, R. (2023). How to Measure a Teacher: The Influence of Test and Nontest Value-Added on Long-Run Student Outcomes. CALDER Working Paper No. 270-0423-2.

Bilan, Y., Mishchuk, H., Roshchyk, I., & Joshi, O. (2020). Hiring and retaining skilled employees in SMEs: problems in human resource practices and links with organizational success. Business: *Theory and Practice*, 21(2), 780-791. https://doi.org/10.3846/btp.2020.12750

Bleiberg, Joshua, and Matthew A. Kraft. (2022). What Happened to the K-12 Education Labor Market During COVID? The Acute Need for Better Data Systems. (EdWorkingPaper: 22-544). Retrieved from Annenberg Institute at Brown University:

https://doi.org/10.26300/2xw0-v642

Bloom, N., & Van Reenen, J. (2011). Human resource management and productivity. In Handbook of labor economics (Vol. 4, pp. 1697-1767). Elsevier. https://doi.org/10.1016/S0169-7218(11)02417-8

Bartanen, Brendan, Aliza N. Husain, David D. Liebowitz, and Laura K. Rogers. (2024). The Returns to Experience for School Principals. (EdWorkingPaper: 24-978). https://doi.org/10.26300/qq95-3q14

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

Burkhauser, S. (2017). How Much Do School Principals Matter When It Comes to Teacher Working

Chi, O. L., & Lenard, M. A. (2023). Can a Commercial Screening Tool Help Select Better Teachers? *Educational Evaluation and Policy Analysis*, *45*(3), 530–539. https://doi.org/10.3102/01623737221131547

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*, 15(3), 510–531. https://doi.org/10.1080/19345747.2021.2018747

DeFeo, D. J., Trang, T., Hirshberg, D., Cope, D., & Cravez, P. (2017). *The Cost of Teacher Turnover in Alaska*. http://hdl.handle.net/11122/7815

Goldhaber, D. & Grout, C. (2024). How Predictive of Teacher Retention Are Ratings of Applicants from Professional References?. CALDER Working Paper No. 296-0324

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., Grout, C., & Wolff, M. (2024). How Well Do Professional Reference Ratings Predict Teacher Performance? *Education Finance and Policy*, 1–41. https://doi.org/10.1162/edfp_a_00421

Goldhaber, D., Ronfeldt, M., Cowan, J., Gratz, T., Bardelli, E., & Truwit, M. (2022). Room for Improvement? Mentor Teachers and the Evolution of Teacher Preservice Clinical Evaluations. *American Educational Research Journal*, 59(5), 1011–1048. https://doi.org/10.3102/00028312211066867

Heneman, H. G., and Judge, T. A. 2003. *Staffing Organizations*. 4th ed. Middleton, WI: McGraw-Hill/Mendota House.

Hoffman, M. and Stanton, C. T. (2024). People, Practices, and Productivity: A Review of New Advances in Personnel Economics. NBER Working Paper No. 32849. https://doi.org/10.3386/w32849

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/10.1016/j.jpubeco.2018.08.011

James, J., Kraft, M. A., & Papay, J. P. (2023). Local supply, temporal dynamics, and unrealized potential in teacher hiring. *Journal of Policy Analysis and Management*, 42(4), 1010–1044. https://doi.org/10.1002/pam.22496

Jovanovic, B. (1979). Job Matching and the Theory of Turnover. *Journal of Political Economy,* 87(5), Part 1. https://doi.org/10.1086/260808

Judge, T. A., Jackson, C. L., Shaw, J. C., Scott, B. A., & Rich, B. L. (2007). Self-efficacy and work-related performance: The integral role of individual differences. *Journal of Applied Psychology*, 92(1), 107–127. https://doi.org/10.1037/0021-9010.92.1.107

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting The Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. *Educational Researcher*, 46(5), 234–249. https://doi.org/10.3102/0013189X17718797

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

Liu, E., & Johnson, S. M. (2006). New Teachers' Experiences of Hiring: Late, Rushed, and Information-Poor. *Educational Administration Quarterly*, 42(3), 324–360. https://doi.org/10.1177/0013161X05282610

Loeb, S., Kalogrides, D., & Beteille, T. (2012). Effective Schools: Teacher Hiring, Assignment, Development, and Retention. *Education Finance and Policy*, 7(3), 269–304. https://doi.org/https://doi.org/10.1162/EDFP_a_00068

Martinková, P., Goldhaber, D., & Erosheva, E. (2018). Disparities in ratings of internal and external applicants: A case for model-based inter-rater reliability. *PLoS ONE*, 13(10), 1–17. https://doi.org/10.1371/journal.pone.0203002

Montgomery, J. D. (1991). Social Networks and Labor-Market Outcomes: Toward an Economic Analysis. *The American Economic Review,*81(5), 1408-1418. https://www.jstor.org/stable/2006929

National Council on Teacher Quality. 2014. "NCTQ Teacher Contract Database: NCTQ District Policy."

Nguyen, T. D., Pham, L. D., Crouch, M., & Springer, M. G. (2020). The correlates of teacher turnover: An updated and expanded Meta-analysis of the literature. *Educational Research Review*, *31*, 100355. https://doi.org/10.1016/j.edurev.2020.100355

Nguyen, Tuan D., Chanh B. Lam, and Paul Bruno. (2022). Is there a national teacher shortage? A systematic examination of reports of teacher shortages in the United States. (EdWorkingPaper: 22-631). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/76eq-hj32

Oyer, Paul, and Scott Schaefer. 2011. "Personnel Economics: Hiring and Incentives." In *Handbook of Labor Economics*, edited by David Car and Orley Ashenfelter, Volume 4, 4:1769–1823. Elsevier B.V. https://doi.org/10.1016/S0169-7218(11)02418-X

Papay, J. P., & Kraft, M. A. (2016). The Productivity Costs of Inefficient Hiring Practices: Evidence From Late Teacher Hiring. *Journal of Policy Analysis and Management*, 35(4), 791–817. https://doi.org/10.1002/pam.21930

Podolsky, A., Kini, T., Darling-Hammond, L., & Bishop, J. (2019). Strategies for Attracting and Retaining Educators: What Does the Evidence Say? *Education Policy Analysis Archives*, 27, 1–47. https://doi.org/10.14507/epaa.27.3722

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., Staiger, D. O., Kane, T. J., & Taylor, E. S. (2012). Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools. *American Economic Review*, 102(7), 3184–3213.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How Teacher Turnover Harms Student Achievement. *American Educational Research Journal*, *50*(1), 4–36. http://aer.sagepub.com.offcampus.lib.washington.edu/content/50/1/4.full.pdf+html

Rucinski, M. (2023). The Effects of Economic Conditions on the Labor Market for Teachers. (EdWorkingPaper: 23-884). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/p3wh-dz18

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting Meta-Analytic Estimates of Validity in Personnel Selection: Addressing Systematic Overcorrection for Restriction of Range. *Journal of Applied Psychology*. https://doi.org/10.1037/apl0000994

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, Jesus F. 2001. "Personnel Selection Methods." In *Personnel Psychology and Human Resource Management: A Reader for Students and Practitioners*, edited by Ivan T Robertson and Cary L Cooper, 1–54. Manchester, UK: John Wiley & Sons, LTD.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, 24(3), 97–118. https://doi.org/10.1257/jep.24.3.97

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

Tables and Figures

Table 1. Candidates and Ratings by Hiring Status and Links to Teacher Outcomes

	All	School-level Screening	Hired	TPEP	Retention
	(1)	(2)	(3)	(4)	(5)
Panel A: Observation Levels					
Applicants	2,501	1,759	444	398	428
Applications	11,268	4,334	448	401	431
Rating-application links	32,812	13,127	1,328	1,203	1,273
Panel B: Reference Ratings					
Overall Criterion					
No Basis for Judgment	0.006	0.007	0.005	0.005	0.005
Below Average	0.017	0.009	0.005	0.005	0.005
Average	0.108	0.101	0.066	0.069	0.067
Very Good	0.179	0.177	0.150	0.147	0.150
Excellent	0.235	0.242	0.218	0.224	0.224
Outstanding	0.297	0.301	0.331	0.327	0.326
Among the Best	0.157	0.163	0.224	0.223	0.223
Observations (ratings-applications)	32,812	13,127	1,328	1,203	1,273

Notes: The columns indicate candidates' progression to school-level screening (used to determine which applicants to interview in person), being hired, and whether a hired applicant is linked to comprehensive performance evaluation scores (from Washington State's Teacher/Principal Evaluation Program) or retention outcomes (from the S-275 Personnel Report maintained by the Washington State OSPI). Applicants refers to unique applicants, where applicants who apply for positions in different hiring years are treated as distinct. Applications refers to job applications to specific positions. In several cases, we observed candidates with a status of *hired* for multiple positions, resulting in a larger number of hires at the application level than the candidate level. The distribution of reference ratings on the *Overall* criterion is presented here. The distributions for the six individual criteria are presented in Figure A2 and Table A2 in the Appendix.

Table 2. Applicant Characteristics and Ratings by Treatment Status

	All	Withheld	Provided	p-value
	(1)	(2)	(3)	(4)
Applicant Characteristics				
Experience	6.008	6.366	5.650	0.224
Advanced Degree	0.448	0.477	0.418	0.176
Observations (applicants)	525	262	263	
Internal Applicant	0.088	0.082	0.093	0.336
Observations (applicants)	2,501	1,253	1,248	
Ratings on *Overall* Criterion				
No Basis	0.006	0.007	0.005	0.319
Below Average	0.008	0.010	0.005	0.016
Average	0.074	0.082	0.066	0.013
Very Good	0.164	0.166	0.163	0.726
Excellent	0.236	0.231	0.241	0.337
Outstanding	0.320	0.311	0.329	0.100
Among the Best	0.192	0.194	0.191	0.788
Observations (ratings)	6,966	3,447	3,519	

Notes: The p-values in column 4 are obtained from paired t-tests of the differences in the means of characteristics and ratings for applicants with withheld and provided reference ratings information. Applicant characteristics are reported at the candidate level, where applicants who apply for positions in different hiring years are treated as distinct. Applicant experience and degree level are self-reported and are only available in the 2018 job year, prior to Spokane's transition to a new ATS provider in 2019. The distribution of ratings is reported at the candidate-rating level, where ratings linked to a candidate in multiple years are treated as distinct.

Table 3. Predicting Advancements in Hiring Process

	Advanced to School-Level		School-level Screening Score		Hired	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Applicant-Job Level						
Reference Rating	0.034**	0.019	0.236***	0.218***	0.031***	0.037***
	(0.016)	(0.012)	(0.035)	(0.032)	(0.008)	(0.008)
Provided Indicator	-0.003	0.001	0.048	0.046	0.015*	0.010
	(0.018)	(0.012)	(0.037)	(0.034)	(0.009)	(0.009)
Provided*Reference Rating	0.003	-0.001	0.001	0.009	-0.011	-0.009
	(0.024)	(0.017)	(0.050)	(0.048)	(0.012)	(0.013)
Job Fixed Effect		Yes		Yes		Yes
Observations	11,271	11,271	4,337	4,337	4,337	4,337
Clusters (applicants)	2,501	2,501	1,759	1,759	1,759	1,759
R-squared	0.003	0.382	0.034	0.359	0.006	0.167
Panel B: Applicant Level						
Reference Rating	-0.027	-0.016	0.264***	0.263***	0.048**	0.060***
	(0.017)	(0.015)	(0.041)	(0.040)	(0.019)	(0.019)
Provided	0.016	0.016	0.049	0.048	0.023	0.020
	(0.018)	(0.017)	(0.044)	(0.043)	(0.020)	(0.020)
Provided*Reference Rating	0.029	0.036	0.016	-0.035	-0.046*	-0.057**
	(0.024)	(0.022)	(0.059)	(0.057)	(0.027)	(0.027)
Competition Controls		Yes		Yes		Yes
Observations	2,501	2,501	1,759	1,759	1,759	1,759
R-squared	0.013	0.177	0.053	0.126	0.010	0.063

Notes: Advanced to School-level Screening is an indicator that an applicant was screened for the job in question. Hired is an indicator that an applicant was hired for the job in question. Screening Score is the school level screening score (standardized ~(0,1)). Reference rating is the summative measure (standardized ~(0,1)) derived from the estimation of a graded response model as described in Section 3.3, averaged at the applicant-job level. Each model in Panel A includes an indicator for whether the applicant is an internal candidate. The control variables in Panel B are an indicator for whether an applicant is an internal candidate, measures of the average quantity and quality of applicants competing for the same jobs, the number of jobs applied to, and the number of ratings associated with the applicant. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4. Accounting for Sample Selection in Predicting Teacher Outcomes Using School-Level Screening Scores and Reference Ratings Among Applicants with Withheld Reference ratings at the Applicant-Job Level

	Heckman Model		OLS Model for Comparison
	Selection (marginal effects)	TPEP Outcome	Teacher Outcome
	(1)	(2)	(3)
Panel A -TPEP Evaluations			
School-level screening score	0.078***	0.277	0.147*
	(0.006)	(0.210)	(0.088)
Reference rating (GRM)	0.009	0.416***	0.411***
	(0.007)	(0.098)	(0.101)
Excluded Variables			
Quantity of competition	-0.001***		
	(0.000)		
Quality of competition	-0.042**		
	(0.017)		
Selection Correction			
Inverse Mills Ratio (λ)		0.250	
		(0.363)	
Observations	2,229	179	179
Panel B - School Retention			
School-level screening score	0.081***	0.074	0.062
	(0.006)	(0.085)	(0.046)
Reference rating (GRM)	0.010	0.043	0.043
	(0.007)	(0.053)	(0.054)
Excluded Variables			
Quantity of competition	-0.001***		
	(0.000)		
Quality of competition	-0.043**		
	(0.018)		
Selection Correction			
Inverse Mills Ratio (λ)		0.023	
		(0.140)	
Observations	2,229	185	185

Notes: The outcome in Panel A is the summative TPEP measure derived in Section 3.4. The outcome in Panel B is an indicator equal to one if the hired applicant is retained in the same school in year $t + 1$. Reference rating is the summative measure derived in Section 3.3. Each model includes an indicator for whether than applicant is an internal candidate, a categorical variable indicating whether the position is for a grade teacher, English language arts, STEM, special education, or other, and school year indicators. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 5. Ranking Applicants for Each Job Using Screening Scores, Reference Ratings, and a Weighted Measure of Screening Scores and Reference Ratings

How often is the same applicant identified as the top-ranked candidate under different rankings scenarios?	Percent Agreement
Screening Score and Reference Rating	38%
Screening Score and Weighted Measure	63%
How often is the top ranked applicant hired?	
Screening Score	55%
Reference Rating	30%
Weighted Measure	42%
Jobs	302
Applications/Applicants	2,860/1,419

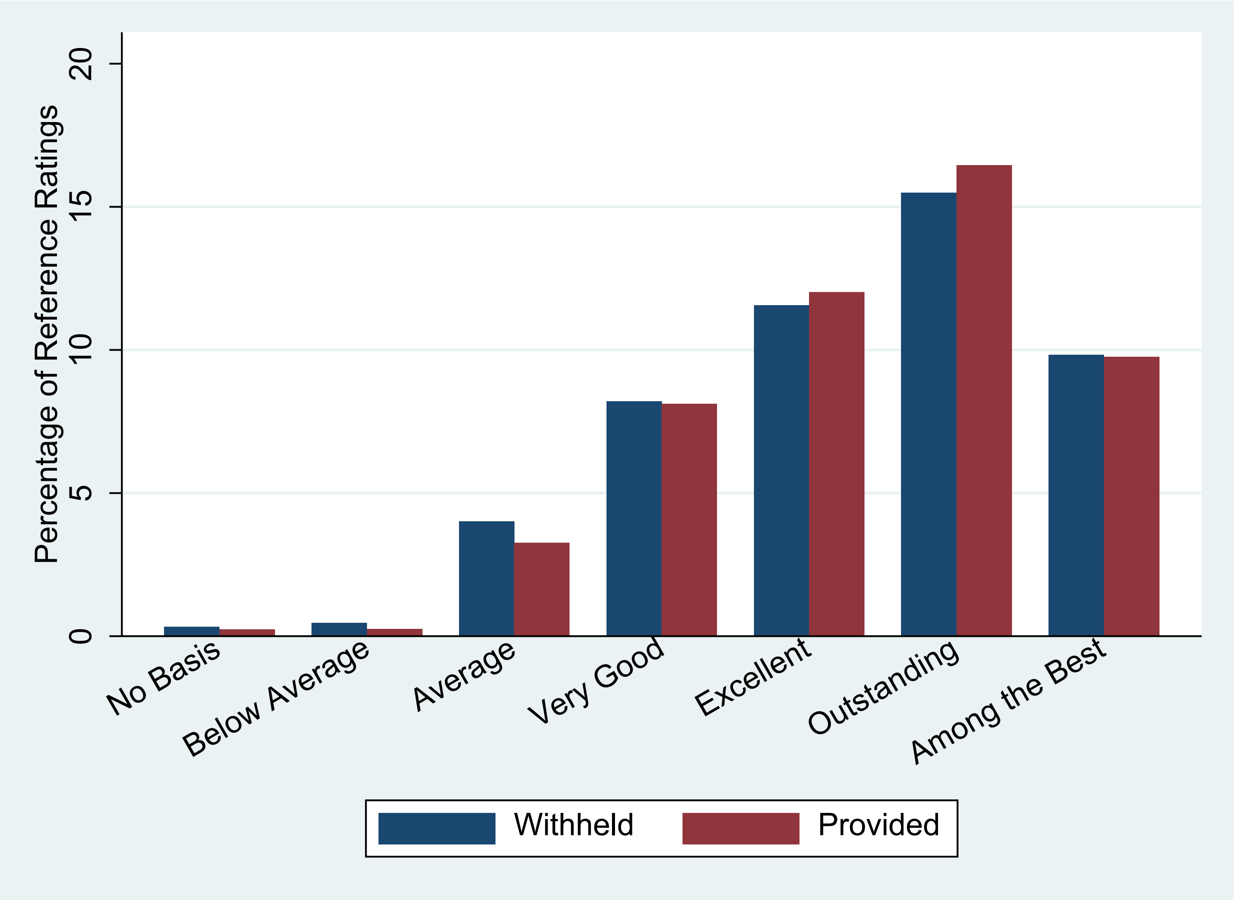
Notes: Screening Score is school level screening score, Reference Rating is the average of the summative reference ratings measures associated with each applicant-job combination, and Weighted Measure is a weighted average of the two using the coefficient estimates from Panel A, column 2 of Table 4 as weights.

Table 6. Predicted Performance Under Alternative Hiring Scenarios Using School-level Screening Score and Reference Ratings Information

Selection Criterion	Mean Predicted TPEP	Obs
Actual hires	0.432	387
Top ranked on:		
Screening Score	0.531	384
Reference Rating	0.514	385
Weighted Rating	0.642	384

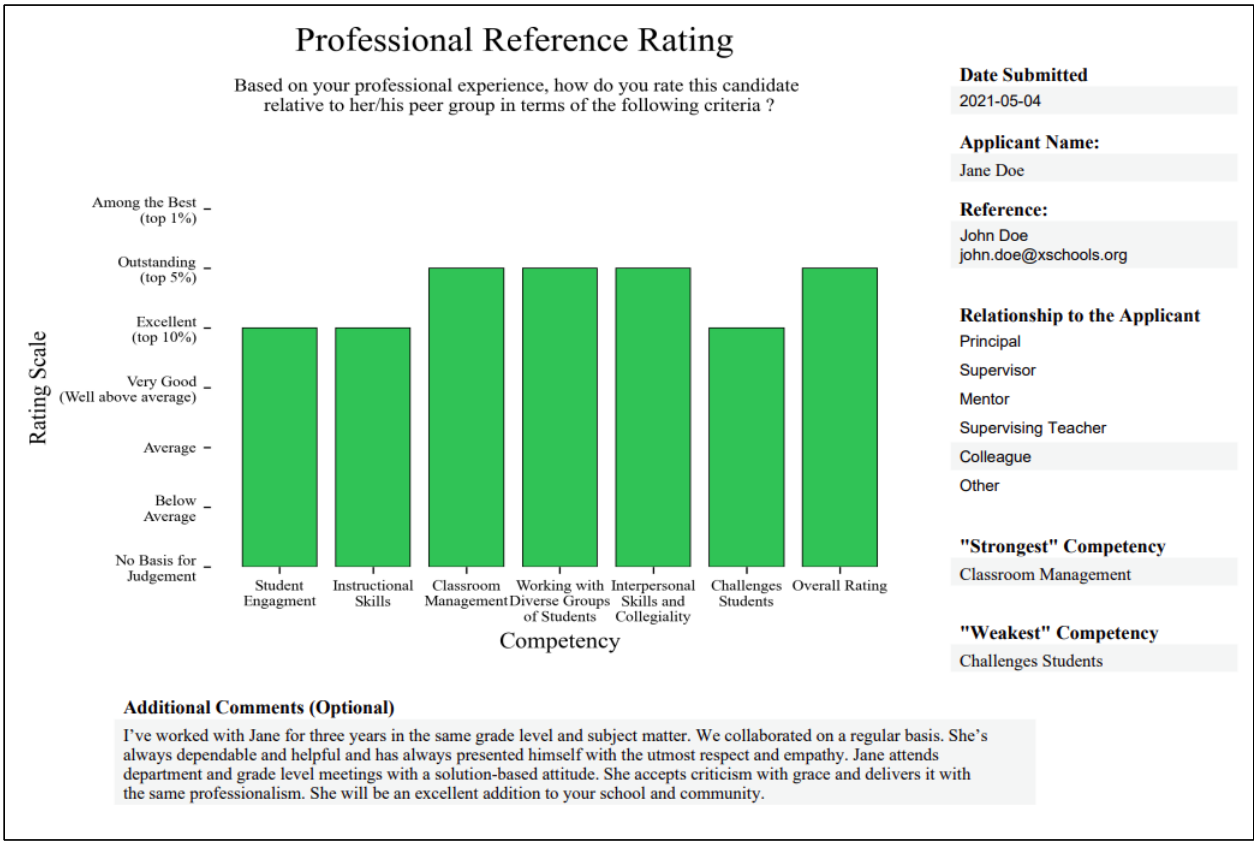
Notes: Screening Score is school level screening score, Reference Rating is the average of the summative reference ratings measures associated with each applicant-job combination, and Weighted Measure is a weighted average of the two using the coefficient estimates from Panel A, column 2 of **Table 4** as weights. Predicted TPEP is calculated following the estimation of the Heckman model.

Figure 1. Distribution of Reference Ratings on the Overall Criterion.



Notes: The distribution of ratings is reported at the applicant-rating level, where ratings linked to an applicant in multiple years are treated as distinct (N = 6,612). Distributions of the individual ratings criteria are presented in Figure A2 in the Appendix.

Figure 2. Example of Reference Ratings Output



Professional Reference Rating

Based on your professional experience, how do you rate this candidate relative to her/his peer group in terms of the following criteria ?

Date Submitted

2021-05-04

Applicant Name:

Jane Doe

Reference:

John Doe
john.doe@xschools.org

Relationship to the Applicant

Principal
Supervisor
Mentor
Supervising Teacher
[✓ Colleague]
Other

"Strongest" Competency

Classroom Management

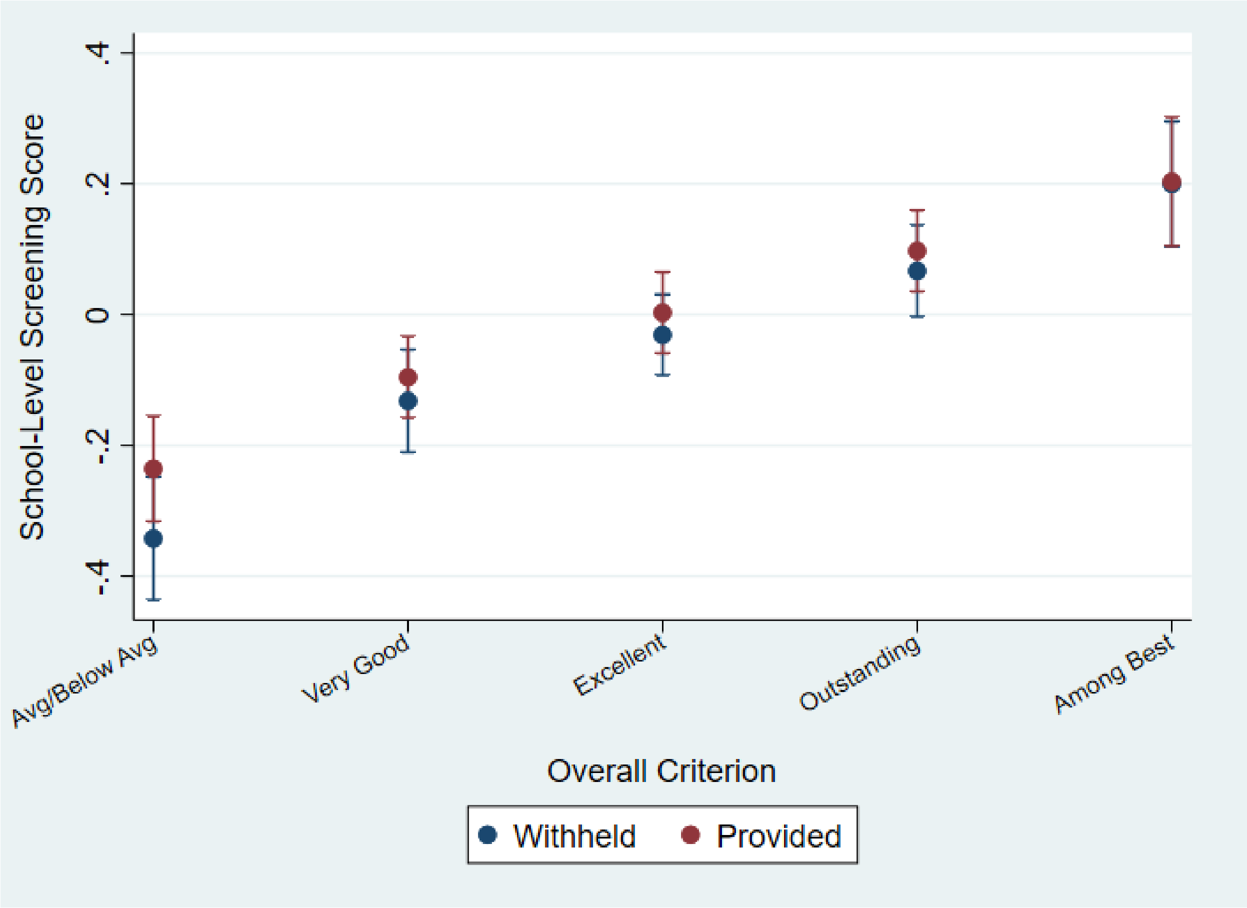
"Weakest" Competency

Challenges Students

Additional Comments (Optional)

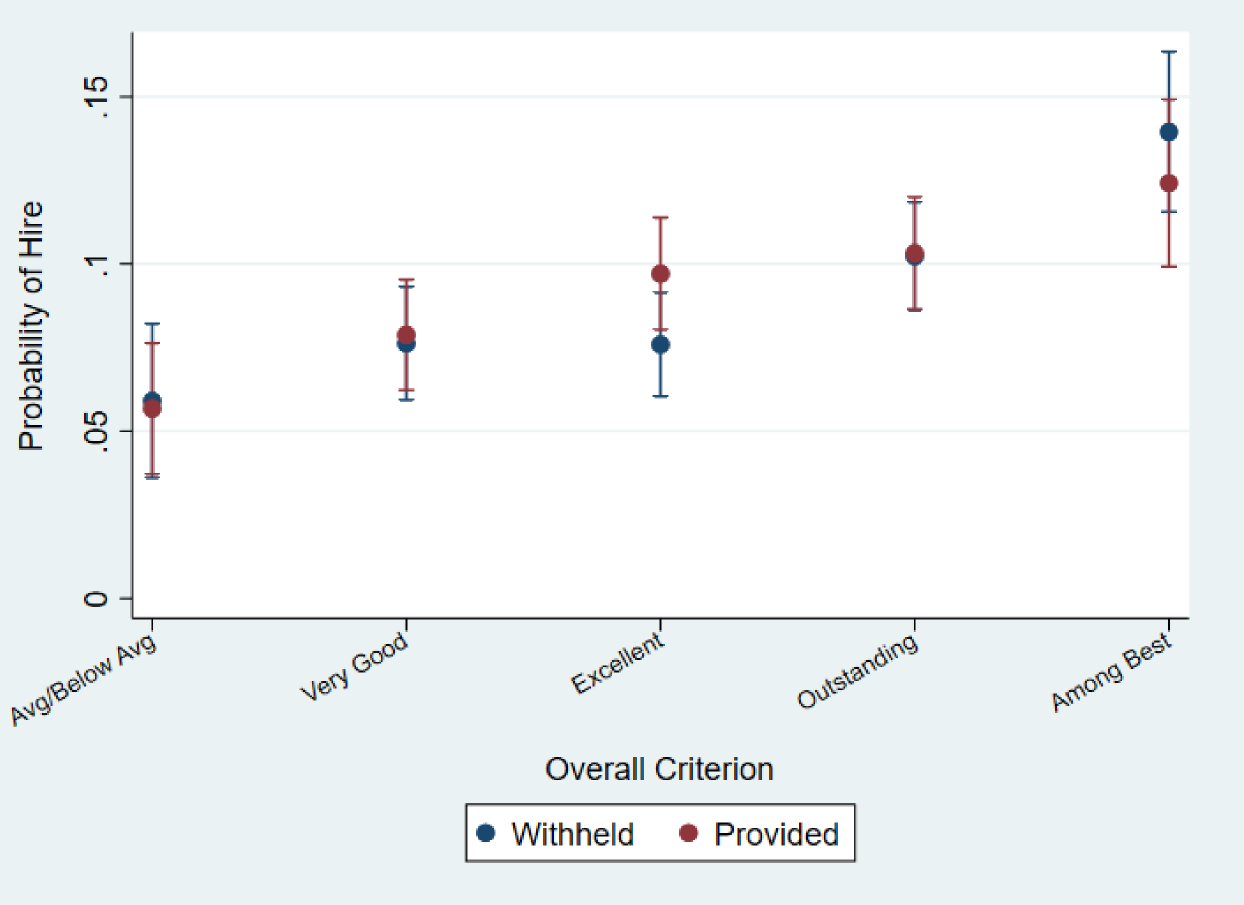
I've worked with Jane for three years in the same grade level and subject matter. We collaborated on a regular basis. She's always dependable and helpful and has always presented himself with the utmost respect and empathy. Jane attends department and grade level meetings with a solution-based attitude. She accepts criticism with grace and delivers it with the same professionalism. She will be an excellent addition to your school and community.

Figure 3. Predicting School-Level Screening Scores Using Categorical Ratings on the Overall Criterion at the Applicant-Job-Ratings Level



Notes: Dependent variable is school-level screening score. The bottom two ratings categories of *Average* and *Below Average* are combined due to small cell sizes. The predictions are generated following the estimation of a linear regression model that includes controls for reference type (e.g., principal, colleague, or university supervisor), an indicator for whether the applicant is an internal candidate, and job fixed effects. The vertical lines represent 95% confidence intervals around the point estimates.

Figure 4. Predicting Hiring Outcomes Using Categorical Ratings on the Overall Criterion at the Applicant-Job-Ratings Level



Notes: Dependent variable is indicator equal to 1 if an applicant is hired. The bottom two ratings categories of *Average* and *Below Average* are combined due to small cell sizes. The predictions are generated following the estimation of a linear probability model that includes controls for reference type (e.g., principal, colleague, or university supervisor), an indicator for whether the applicant is an internal candidate, and job fixed effects. The model is estimated conditional on the application being advanced to the school-level screening stage of the hiring process. The vertical lines represent 95% confidence intervals around the point estimates.

Appendix: Supplemental Tables and Figures

Table A1. 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

Table A2. Distribution of Individual Ratings Criteria by Treatment Status

Criterion	Rating	Withheld	Provided	p-value
Student Engagement	No Basis	0.032	0.027	0.201
	Below Average	0.010	0.005	0.043
	Average	0.068	0.062	0.292
	Very Good	0.159	0.144	0.083
	Excellent	0.209	0.226	0.085
	Outstanding	0.297	0.310	0.229
	Among the Best	0.225	0.225	0.982
Instructional Skills	No Basis	0.036	0.040	0.374
	Below Average	0.008	0.005	0.121
	Average	0.080	0.066	0.026
	Very Good	0.172	0.159	0.130
	Excellent	0.226	0.240	0.181
	Outstanding	0.283	0.294	0.299
	Among the Best	0.195	0.196	0.882
Classroom Management	No Basis	0.064	0.057	0.241
	Below Average	0.015	0.012	0.255
	Average	0.102	0.080	0.002
	Very Good	0.158	0.167	0.280
	Excellent	0.217	0.225	0.385
	Outstanding	0.277	0.284	0.491
	Among the Best	0.168	0.174	0.531
Working with Diverse Groups of Students	No Basis	0.033	0.032	0.668
	Below Average	0.005	0.003	0.160
	Average	0.057	0.045	0.017
	Very Good	0.140	0.142	0.868
	Excellent	0.213	0.203	0.288
	Outstanding	0.310	0.325	0.188
	Among the Best	0.240	0.251	0.286
Interpersonal Skills	No Basis	0.004	0.005	0.531
	Below Average	0.010	0.011	0.882
	Average	0.072	0.057	0.010
	Very Good	0.131	0.134	0.739
	Excellent	0.223	0.217	0.528
	Outstanding	0.306	0.322	0.145
	Among the Best	0.252	0.253	0.916
Challenges Students	No Basis	0.042	0.035	0.138
	Below Average	0.009	0.006	0.188
	Average	0.084	0.072	0.051
	Very Good	0.164	0.150	0.112
	Excellent	0.227	0.251	0.017
	Outstanding	0.291	0.295	0.754
	Among the Best	0.180	0.188	0.391
Observations		3,447	3,519	

Notes: The distribution of ratings is reported at the applicant-rating level, where ratings linked to an applicant in multiple years are treated as distinct. The p-values are obtained from paired t-tests of the differences in the means of characteristics and ratings for applicant with withheld versus provided reference ratings.

Table A3. Distribution of Competencies Identified as Strongest/Weakest by Treatment Status

	Strongest Competency			Weakest Competency		
Criterion	Withheld	Provided	p-value	Withheld	Provided	p-value
Challenges Students	0.058	0.055	0.675	0.252	0.255	0.831
Diverse Groups	0.269	0.267	0.886	0.129	0.143	0.099
Student Engagement	0.200	0.223	0.023	0.059	0.057	0.752
Instructional Skills	0.173	0.164	0.319	0.108	0.101	0.299
Interpersonal Skills	0.216	0.201	0.132	0.186	0.194	0.370
Classroom Management	0.084	0.089	0.448	0.265	0.251	0.158
Observations	3,447	3,519		3,447	3,519	

Notes: The distribution of ratings is reported at the candidate-rating level, where ratings linked to an applicant in multiple years are treated as distinct. The p-values are obtained from paired t-tests of the differences in the means of characteristics and ratings for applicant with withheld versus provided reference ratings.

Table A4. Predicting Advancement in Hiring Process at the Applicant-Job-Reference Rating Level

	Advanced to School-level Screening		School-level Screening Score		Hired	
	(1)	(2)	(3)	(4)	(5)	(6)
Reference Rating	0.001	0.005	0.152***	0.140***	0.021***	0.020***
	(0.011)	(0.007)	(0.022)	(0.019)	(0.005)	(0.004)
Provided	-0.011	-0.003	0.017	0.037	0.010	0.003
	(0.021)	(0.014)	(0.041)	(0.036)	(0.010)	(0.009)
Provided*Reference Rating	0.005	0.003	-0.017	-0.015	-0.007	-0.003
	(0.016)	(0.010)	(0.030)	(0.026)	(0.007)	(0.006)
Job Fixed Effect		Yes		Yes		Yes
Observations	32,812	32,812	13,127	13,127	13,127	13,127
Clusters (applicants)	2,501	2,501	1,759	1,759	1,759	1,759
R-squared	0.007	0.390	0.024	0.381	0.004	0.186

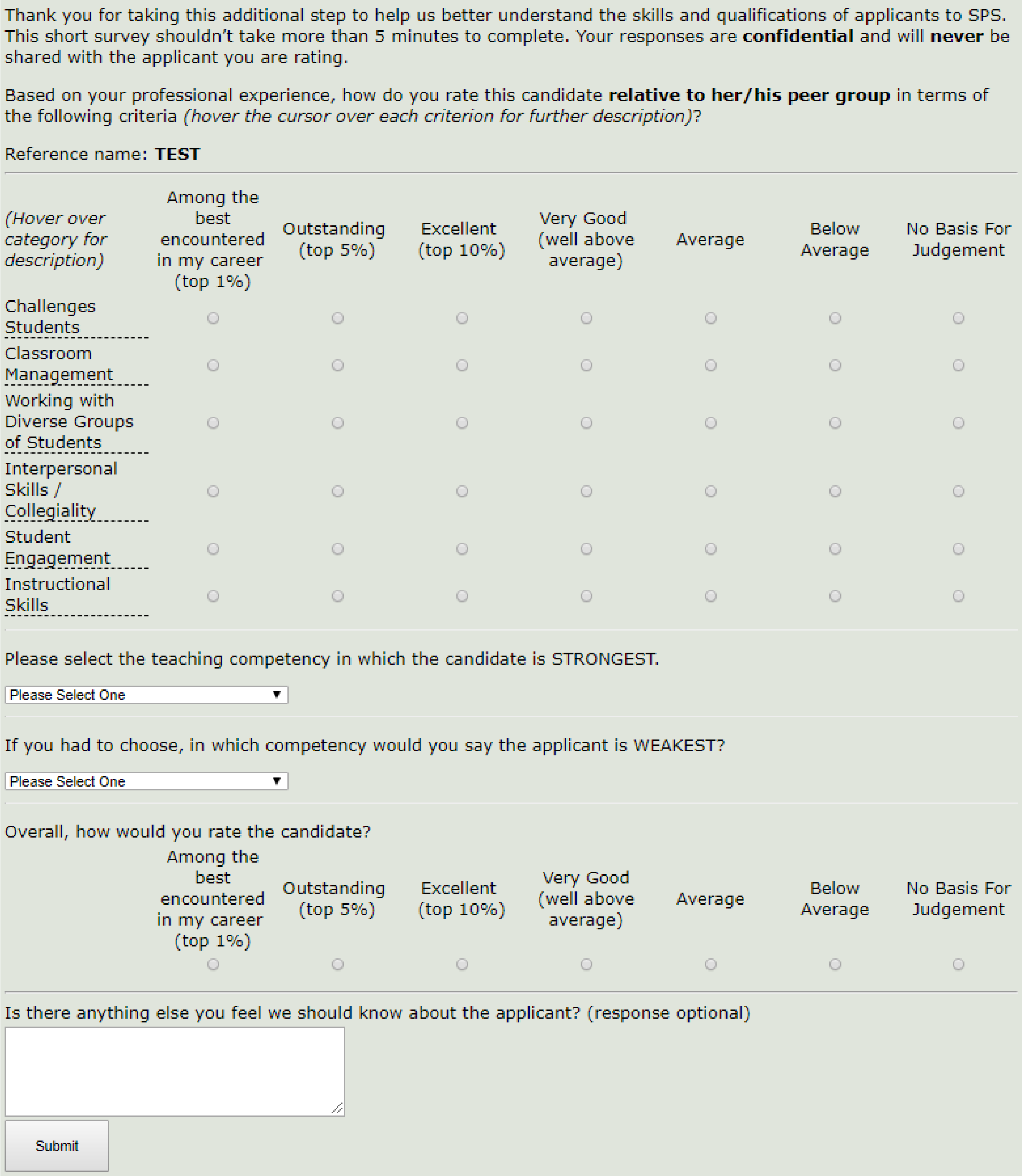
Notes: Advanced to School-level Screening is an indicator that an applicant was screened for at least one job. Hired is an indicator that an applicant was hired for a job. Screening Score is the average of the school-level screening scores associated with that applicant (standardized $\sim(0,1)$ before averaging). Reference rating is the summative measure (standardized $\sim(0,1)$) derived from the estimation of a graded response model as described in Section 3.3, averaged at the applicant-job level. Each model includes an indicator for whether the applicant is an internal candidate and a categorical variable indicating whether the references relationship to the applicant is Principal/Other Supervisor, Colleague, Cooperating Teacher, University Supervisor, or Other. * p < 0.10, ** p < 0.05, *** p < 0.01

Table A5. Accounting for Sample Selection in Predicting Retention Outcomes Using School-Level Screening Scores and Reference Ratings Among Applicants with Withheld Reference ratings at the Applicant-Job Level

	Heckman Model		Probit Model
	Selection	School Retention	School Retention
	(1)	(2)	(3)
School-level screening score	0.721***	0.208	0.187
	(0.046)	(0.262)	(0.129)
Reference rating (GRM)	0.093	0.127	0.126
	(0.059)	(0.146)	(0.146)
Excluded Variables			
Quantity of competition	-0.010***		
	(0.003)		
Quality of competition	-0.383**		
	(0.159)		
Observations	2,229	185	185

Notes: School-level screening score is (standardized ~(0,1)). Reference rating is the summative measure (standardized ~(0,1)) derived from the estimation of a graded response model (GRM) and described in Section 3.3. Each model includes an indicator for whether than applicant is an internal candidate, a categorical variable indicating whether the position is for a grade teacher, English language arts, STEM, special education, or other, and school year indicators. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure A1. 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.

Please Select One ▼

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

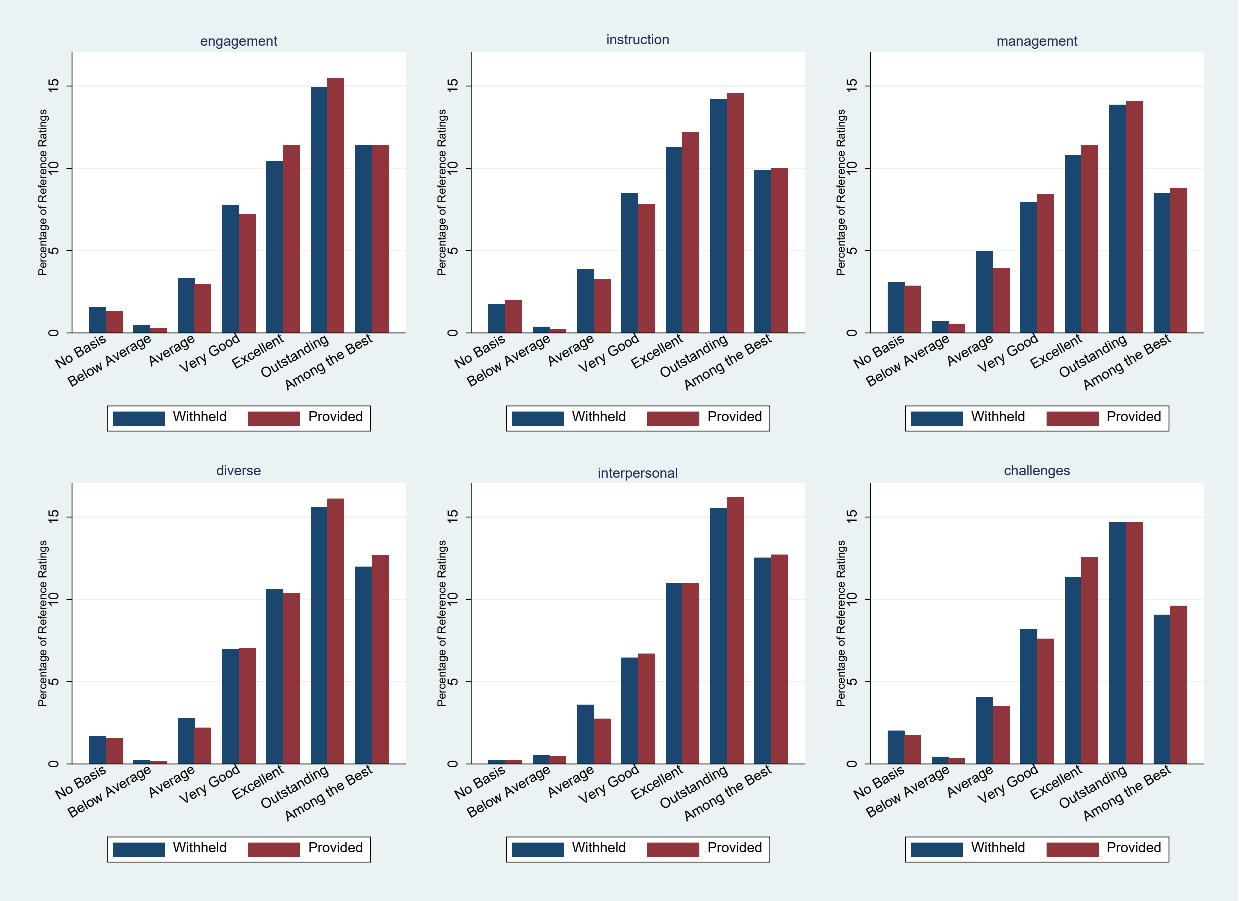
Please Select One ▼

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
○	○	○	○	○	○	○

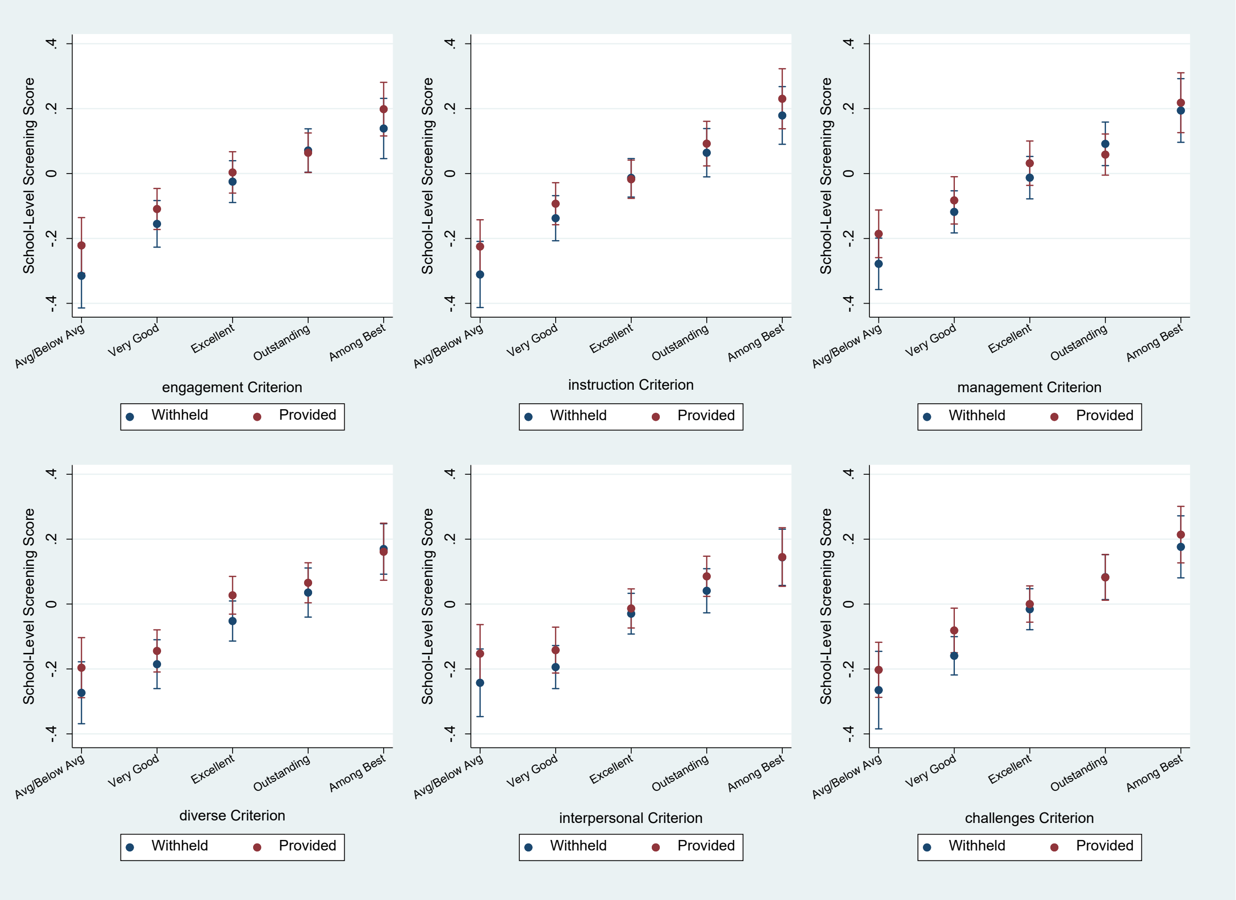
Is there anything else you feel we should know about the applicant? (response optional)

Figure A2. Distribution of Individual Ratings Criteria by Blinded Status



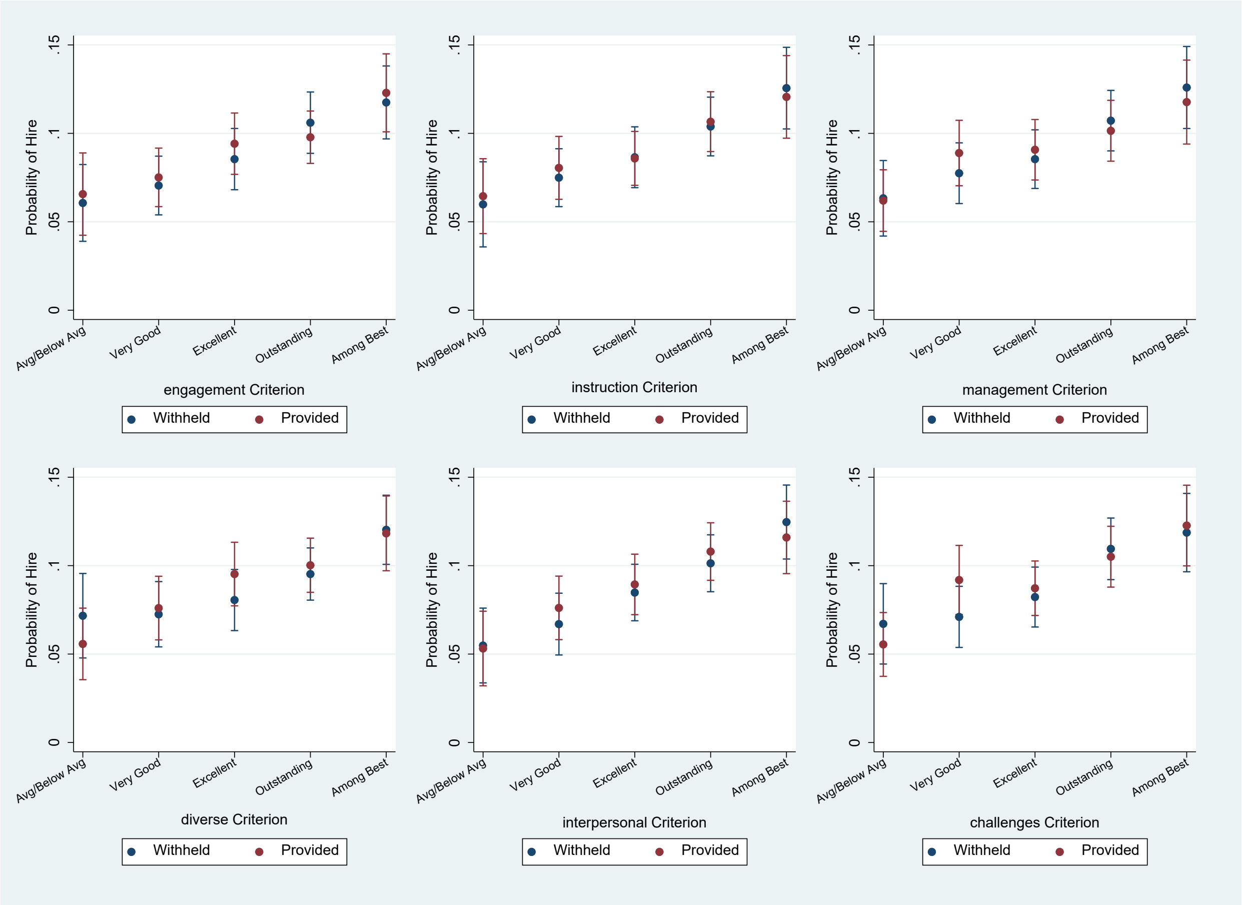
Notes: The distribution of ratings is reported at the applicant-rating level, where ratings linked to an applicant in multiple years are treated as distinct (N = 6,612).

Figure A3. Predicting Screening Scores Using Individual Categorical Ratings Criteria at the Applicant-Job-Ratings Level



Notes: Dependent variable in each plot is school-level screening score. The bottom two ratings categories of *Average* and *Below Average* are combined due to small cell sizes. The predictions are generated following the estimation of a linear regression models that include controls for reference type (e.g., principal, colleague, or university supervisor), an indicator for whether the applicant is an internal candidate, and job fixed effects. The vertical lines represent 95% confidence intervals around the point estimates.

Figure A4. Predicting Hiring Outcomes Using Individual Ratings Criteria at the Applicant-Job-Ratings Level



Notes: Dependent variable in each model is an indicator equal to 1 if an applicant is hired. The bottom two ratings categories of *Average* and *Below Average* are combined due to small cell sizes. The predictions are generated following the estimation of a linear probability models that includes controls for reference type (e.g., principal, colleague, or university supervisor), an indicator for whether the applicant is an internal candidate, and job fixed effects. The models are estimated conditional on the application being advanced to the school-level screening stage of the hiring process. The vertical lines represent 95% confidence intervals around the point estimates.