How Predictive of Teacher Retention Are Ratings of Applicants from Professional References?

Dan Goldhaber
Cyrus Grout

March 2024

WORKING PAPER No. 296-0324

CALDER
National Center for Analysis of
Longitudinal Data in Education Research

AIR®

How Predictive of Teacher Retention Are Ratings of Applicants from Professional References?

Dan Goldhaber
American Institutes for Research / CALDER
University of Washington

Cyrus Grout
Center for Education Data & Research
University of Washington

Contents

Acknowledgments

This work is supported by the Institute of Education Sciences (grant # R305A170060). All opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the institutions to which the authors are affiliated or the study's funder. We thank Spokane Public Schools for partnering with us on this work and James Cowan and Roddy Theobald for insightful comments. Correspondence regarding this article should be addressed to Dan Goldhaber, Center for the Analysis of Longitudinal Data in Education Research, American Institutes for Research, 3876 Bridge Way N, Suite 201, Seattle, WA 98103. Email: dgoldhaber@air.org • www.caldercenter.org.

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

How Predictive of Teacher Retention Are Ratings of Applicants from Professional References?

Dan Goldhaber & Cyrus Grout

CALDER Working Paper No. 296-0324
March 2024

Abstract

Turnover in the teacher workforce imposes significant costs to schools, both in terms of student achievement and the time and expense required to recruit and train new staff. This paper examines the potential for structured ratings of teacher applicants, solicited from their professional references, to inform hiring decisions through the selection of teachers who are less likely to turn over. Specifically, we analyze the predictive validity of reference ratings with respect to retention outcomes among subsequently employed applicants. We find that a summative reference ratings measure is modestly predictive of retention in a teacher's school, with a one-standard deviation change associated with a 3.2-percentage point increase in the probability of school retention. When we account for rater fixed effects, we find substantially stronger relationships between reference ratings and retention, with a one-standard deviation change in our summative ratings measure associated with an increase in the probability of school retention of 8.5 percentage points. These findings suggest that raters themselves are a large source of variation in the distribution of reference ratings. So, while we find predictive validity of professional ratings, their potential to inform good hiring decisions depends on, among other things, the ability of hiring managers to account for rater variation when interpreting references' assessments of applicants.

1. Introduction

The potential to reduce employee turnover in firms by making better hiring decisions has been explored in the fields of personnel economics and industrial psychology (Heneman & Judge, 2003; Shaw & Lazear, 2007). Reducing employee turnover is also a matter of concern in public schools. Interviews with principals, for example, reveal their concern about retaining newly hired teachers (Harris et al., 2010), and rightly so. Attrition among teachers is high in their early years in the profession. For example, in the first five years of teaching—the period in which teachers typically see significant improvement (Papay & Kraft, 2015; Rockoff, 2004)—40% to 50% leave the profession (Ingersoll & Strong, 2011). Unsurprisingly, a range of evidence suggests that teacher turnover negatively impacts student achievement as well as teachers who remain on the job. The expenses associated with recruiting, hiring, and training new staff because of teacher turnover also impose financial costs on school systems (Barnes et al., 2007; DeFeo et al., 2017). Given these costs, the far-reaching impact teachers have on student success (Chetty et al., 2014), and the sheer size of the teaching profession [with over 3 million members, teaching is the largest public-sector occupation in the United Sates (National Center for Education Statistics, 2022)], understanding which hires are likely to stay or leave is an important issue for practice and research. However, to date, few studies examine the predictive value of applicant information for teacher outcomes, including turnover (Bruno & Strunk, 2019; Chi & Lenard, 2022; Goldhaber et al., 2017; Jacob et al., 2018; Sajjadiani et al., 2019).[1]

In this paper, we analyze the extent to which novel, low-cost information about applicants—structured categorical ratings by professional references—predicts teacher retention

[1] For a broader assessment of the factors that influence teacher attrition, see Nguyen et al. (2020) who synthesize findings from 120 empirical analyses of the factors that correlate with teacher retention, updating earlier work by Borman and Dowling (2008).

at the school, district, and workforce levels, and what factors moderate the relationship between reference ratings and retention. The ratings information is relatively low-cost because it builds off a common hiring practice: the collection of letters of recommendation (Salgado, 2001). Recommendation letters are already perceived as valuable sources of information by hiring managers; for example, a survey of principals in North Carolina found that the three most important artifacts in teacher applicants' portfolios were the candidate's resume, references, and letters of recommendation (Nodoye et al., 2012). Goldhaber et al. (2017) also found that letters of recommendation heavily inform principals' ratings of applicants. To the extent that quantifiable ratings from professional references provide additional useful information, they have the added benefit of taking less time to collect and review and may allow for easier comparisons between candidates. To see if professional reference ratings of applicants provide information about whether job applicants will remain on their job (contingent on being hired), we analyze applicant ratings from professional references in Spokane Public Schools (SPS) between June 2015 and October 2018. We link these reference ratings to teacher retention outcomes for teacher applicants who were observed in the workforce in Washington state (in SPS or other districts) in the 2015-16 through 2019-20 school years.

When professional references rate job applicants, we find that these ratings are predictive of school-level retention among classroom teachers hired into a new position, with a one-standard deviation increase in a summative applicant ratings measure associated with a three-percentage point increase in retention. We do not consistently find statistically significant evidence that reference ratings are predictive of retention at the district and state levels; here findings are sensitive to model specification. When rater-fixed effects are included in the model (so comparisons are between teacher applicants *within* rater), we find that reference ratings are

significantly predictive of retention at the school, district, and state levels. One implication of our findings is that school districts need to wrestle with interrater reliability of professional references to make applicant ratings more useful.

2. Teacher Applicant Information and Retention

While there is a significant literature on in-service factors that predict teacher turnover and attrition (Nguyen et al., 2020), there is only limited evidence about the extent to which information about teacher applicants—particularly those new to the profession—might be used to hire applicants with a relatively low propensity for attrition.[2] The evidence that does exist suggests that when *school districts* score or rate applicants, those ratings can help predict the likelihood an applicant will stay in the job if hired. For example, Goldhaber et al. (2017) analyzed teacher applicant screening scores generated by school principals in the process of determining which applicants to interview in person. Their study found that a one-standard deviation increase in an applicant's screening score was associated with a three-percentage point decrease in the propensity to leave the district the following year (the baseline level of attrition was about 16%). They also found that screening scores were more predictive of attrition for teachers with more experience compared to teachers who were new to the profession.

Bruno and Strunk (2019) analyzed a centralized district screening process in Los Angeles Unified School District that used district-developed rubrics to score applicants on a series of criteria, including subject area preparation, written responses to student-related scenarios, sample lessons, and structured interviews. They found that a one-standard deviation increase in the composite screening score was associated with a 1.6-percentage point decrease in the probability of school-level turnover (scores were associated with a lower probability of leaving the district

[2] Most of these studies also assessed the extent to which applicant information is predictive of measures of in-service teacher performance.

but the estimates were not statistically significant). Jacob et al.'s (2018) study of a multi-stage screening process used by Washington DC Public Schools (DCPS)—which included written assessments of pedagogy and content knowledge, personal interviews, and teaching auditions—found that a one-standard deviation increase in screening scores predicted a 4.4-percentage point decrease in the probability of attrition from the hiring school in the next year (scores were not significantly predictive of leaving DCPS). More recently, Chi and Lenard (2023) investigated whether a commercially available screening tool—Frontline Education's TeacherFit instrument—is predictive of teacher retention. The TeacherFit assessment consists of a survey completed by applicants that seeks to assess their attitudes, beliefs, habits and personality traits, and which takes roughly 20-30 minutes to complete. The instrument generates an overall score and scores on six different dimensions characterizing the applicant. In contrast to the above studies which found that teachers who perform better on district assessment measures tended to have higher rates of retention, Chi and Lenard found that a one standard deviation increase in the TeacherFit score was associated with a 3.4-percentage point *decrease* in the propensity to stay in the hiring school in the following year and a 2.4-percentage point decrease in propensity to stay in the district. Finally, Sajjadiani et al. (2019) examine the potential of using machine learning to interpret applicants' work history in the Minneapolis Public School District (an approach that could reduce assessments of applicants after the initial set up). They generated three measures of applicant quality: work experience relevance, tenure history, and attributions for previous turnover. They found that a one standard deviation increase in work experience relevance was associated with an 8% decrease in the hazard of voluntary turnover. Tenure history was found to be predictive of both voluntary and involuntary turnover, with a one standard deviation increase associated with decreases in the hazard of turnover of 11% and 13% respectively. While this

literature suggests the potential of applicant ratings or scores generated by districts as predictors of attrition, it does not address the potential of ratings generated by professional references.

Prior research on ratings from professional references suggests they are predictive of future performance but does not yet address whether they predict attrition. Analyzing the same reference ratings data that is the subject of our current analysis, for example, Goldhaber et al. (2023a) found that the ratings of teacher applicants by their professional references were significantly predictive of subsequent in-service performance evaluations and value-added in math and that these relationships were stronger for some types of references (e.g., those identified as an applicant's *Principal/Other Supervisor* or *Instructional Coach/Department Chair*) and for applicants with at least some prior teaching experience. The authors also found that the estimated relationship between reference ratings and performance was substantially stronger when rater-fixed effects were introduced to account for rater-driven variation in the ratings of applicants. These findings suggest the need to be attentive to the relationship between the applicant and the professional reference, the applicant's level of experience, and *rater*-driven variation in the reference ratings. But the question remains whether these same professional reference ratings tell districts anything about the likelihood a candidate, if hired, would stay or leave.

3. Data and Analytic Sample

3.1 Data and Measures

In this section, we describe the teacher application process in SPS, the collection of reference ratings, measures of applicant quality derived from the reference ratings data, and the administrative data used to characterize teacher mobility.

To apply for a teaching position in SPS, applicants begin by creating a profile in the applicant tracking system (ATS) used by the district. The profile contains information about the

applicant's educational background, credentials, work history, resume, and personal statements. Applicants are also asked to provide contact information for at least three professional references. Newly listed references receive an email from the ATS prompting them to upload a confidential letter of recommendation, which is then appended to the applicant's profile. Once a profile is established, the applicant can apply to specific job openings.[3]

In 2015, SPS began soliciting reference ratings with a slight modification to the letter of recommendation submission process: after submitting a letter, the professional reference was redirected to an online survey form (see Figure A1 in the Appendix). The survey form asked the following question: "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 evaluation criteria are: *Challenges Students*, *Classroom Management*, *Working with Diverse Groups of Students*, *Interpersonal Skills*, *Student Engagement*, and *Instructional Skills* (see Table A1 in the Appendix for descriptions of each criterion). References were also asked to rate the applicant *Overall*. References rate the applicant using the following scale: *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.*

As discussed in Goldhaber et al. (2021), applicants are likely to have good relationships with their professional references; not surprisingly there is a tendency for references to describe applicants positively or engage in "cheerleading." With this in mind, we adopted the relative percentile ranking method and concentrated the ratings categories at the upper end of the distribution. We also pose two questions that are not subject to cheerleading: "Please select the

[3] In SPS, job postings typically refer to a specific position (e.g., Grade 3 Teacher at X Elementary School) and it is common for an applicant to apply to multiple job postings.

competency in which the applicant is *Strongest*"; and "If you had to choose, in which competency would you say the applicant is *Weakest*?"

The ratings for each of the six evaluation criteria and the *Overall* rating are ordered categorical data and we use these categorical measures to model the relationship between reference ratings and retention outcomes. To facilitate analysis, we also construct a continuous summative ratings measure derived from the estimation of a graded response model (GRM). Following Chen et al. (2021), we specify the model represented in equation (1) where the probability of observing rating level k or higher for evaluation criterion c and rating of applicant i by reference j is expressed as:

$$\Pr\left(Y_{ijc} \geq k \middle| GRM_{ij}\right) = \frac{\exp\{a_c(GRM_{ij} - b_{ck})\}}{1+\exp\{a_c(GRM_{ij} - b_{ck})\}}, \quad (1)$$

where a_c represents the discrimination of criterion c, b_{ck} is the kth cutpoint of criterion c, and GRM_{ij} is the latent quality expressed by reference rating ij. We estimate the model with standard errors clustered at the applicant level and use the estimated values GRM_{ij} as a summative measure of applicant quality represented by each reference rating.[4]

We link reference ratings to other information in a job applicant's file using a unique applicant ID. Thus, we also have information on applicant names, certification IDs, and employee IDs, which are used to link the application data to Washington state administrative data. Specifically, we link the reference rating and teacher application data to statewide administrative data from two sources: the Washington State S-275 personnel reporting system and the Comprehensive Education Data and Research System (CEDARS) database. The S-275

[4] Note that we cannot calculate $\widehat{GRM_{ij}}$ in instances where one or more ratings criteria are rated as "No basis for judgement," which is occurs in 14% of the ratings we collected (including ratings of applicants that we do not subsequently observe in the workforce).

data are maintained by the Office of the Superintendent of Public Instruction (OSPI) and report position assignments, compensation, experience, degree level, ethnicity, and gender of certificated and classified staff under contract as of October 1 of each school year. The CEDARS database includes IDs linking teachers to classrooms and schools, and student-level data in CEDARS allows us to construct school-level measures of student characteristics and performance on standardized tests.[5]

3.2 Analytic Sample

Our study sample is anchored by applicants who applied for one or more teaching positions in SPS during the 2015 to 2018 hiring years and for whom we collected one or more reference ratings. We observe teacher mobility outcomes for the subset of applicants who are employed by a public school district in Washington during the 2015-16 to 2019-20 school years. We link reference ratings to applicants if they were generated in the same year or in the year prior. This results in a small percentage of reference ratings being linked to multiple application years; fewer than 5% of ratings in our analytic sample are linked to multiple application years.

We restrict the sample in several ways. First, we restrict the analytic sample to teachers who moved into a new teaching position. We characterize "movers" as classroom teachers who meet one of the following criteria: they are new to the K-12 public school teacher workforce in Washington; they are working in a different school than in the prior year; they moved into a classroom teaching position from a non-classroom teaching position.[6] We exclude non-movers because in applying for a position in SPS, they expressed interest in leaving their existing position and might be expected to maintain that interest in moving in the ensuing school year.

[5] Student information includes data on gender, race/ethnicity, free or reduced-lunch participation, migrant status, homelessness status, and standardized test scores.
[6] We identify classroom teaching positions as those with a duty code of 31 (Elementary Homeroom Teacher), 32 (Secondary Teacher), 33 (Other Teacher), or 34 (Elementary Specialist Teacher) in the state's S-275 personnel data.

Second, we exclude teachers assigned to multiple schools because applying a binary definition of mobility to these teachers is problematic.[7] Finally, we exclude ratings where we cannot estimate the summative ratings measure GRM described above and applicants for whom we do not observe the full set of control variables.[8]

In Table 1, we present descriptive statistics for the 3,532 reference ratings of 1,124 unique applicants in our analytic sample.[9] Statistics for the full sample are presented in column (1). In columns (2) to (4) we calculate the *differences* in means between public school stayers and leavers at the school, district, and the state levels, x: $mean(x|stay) - mean(x|leave)$. Bold text indicates statistically significant differences ($p < 0.05$). Regarding teacher characteristics, the proportion of teachers with no prior experience is 6 points lower among stayers relative to leavers while the proportion of experienced teachers (with 6+ years of experience) is 8 points higher. This is consistent with the literature on teacher retention, which finds lower rates of retention among inexperienced teachers (Nguyen et al. 2022). Otherwise, we do not observe significant differences in teacher characteristics between stayers and leavers. Statistics for teacher and school characteristics are calculated at the teacher level (1,124 observations) and statistics for reference ratings measures are calculated at the ratings level (3,532 observations).

[7] Of the ratings that would have otherwise been included in our analytic sample, 7.8% are associated with teachers assigned to multiple schools.

[8] We are unable to estimate GRM_{ij} for 6.4% of the ratings that would have otherwise been included in our analytic sample. Taken together, the restrictions of being able to estimate GRM and observing the full set of control variables excludes 11.8% of reference ratings. However, a smaller proportion of unique applicants are excluded (6.1%) since many applicants have multiple ratings and we can often estimate GRM for at least one of them.

[9] For the purposes of this study, applicants who apply for a teaching position in different years are treated as distinct applicants/teachers because the characteristics of those applicants/teachers, such as their level of experience and work history, change over time. Among the 1,050 unique individuals in our analytic sample, 145 appear in multiple years.

Overall, we observe a 68% rate of retention at the school level, 78% at the district level, and 86% at the state level. As shown in column (1), the applicants in our analytic sample have varied levels of experience, with 32% of ratings associated with teachers entering their first year of teaching and a similar proportion (28%) associated with teachers who have 6 or more years of teaching experience. The majority of teachers are female (75%) and white (97%), and nearly half hold an advanced degree (47%). Over half of the teachers in our sample teach in districts other than SPS (53%) and a majority are in elementary positions. The teachers in our sample work in schools with above-average percentages of students from low-income households: the average school-level proportion of students eligible for free or reduced-price lunch (FRL) is 57% compared to a state-wide average of 47% in 2018-19.[10] The teachers in our sample also work in schools with below-average levels of student achievement on standardized tests: 14% of a standard deviation and 20% of a standard deviation below average for math and English language arts (ELA), respectively. The below-average levels of student achievement in our analytic sample, which consists of teachers who have moved into a new position, are consistent with research finding higher rates of teacher mobility (and by extension, more hiring) in lower-achieving schools (Goldhaber et al., 2023b).

We observe a number of significant differences between stayers and leavers at the school level (column (2)). The proportion of teachers who work in SPS is nine points lower among stayers. This likely reflects the fact that there are more opportunities for within-district mobility in SPS, which is the largest school district in eastern Washington. Similarly, the proportion of teachers working in elementary schools (which are more numerous than secondary schools), is seven points lower among stayers relative to leavers. The figures are also consistent with lower

[10] See: https://washingtonstatereportcard.ospi.k12.wa.us/ReportCard/ViewSchoolOrDistrict/103300.

rates of school-level retention at higher-poverty and lower-achieving schools. These patterns suggest that it will be important to control for district size and school context when modeling the relationship between reference ratings and retention. These differences are not statistically significant when we look at the differences between stayers and leavers at the district (column (3)) or state (column (4)) levels.

Turning to the reference ratings, we find that very few ratings on the *Overall* criterion (5%) fall in the bottom two ratings categories and that over half of ratings are in the top two categories. That said, the ratings exhibit a good deal of variation with between 14% and 35% falling within each of the top four categories. In comparing the ratings of stayers and leavers we find that the proportion of teachers rated in one of the bottom two ratings categories (*Average* or *Below Average*) on the *Overall* criterion is significantly lower among stayers at the school, district and state levels. Conversely, the proportion of teachers rated in the top category (*Among the best (top 1%)*) is significantly higher (+4 percentage points) among stayers at the school level. The proportions of ratings falling in the middle three categories do not exhibit any significant differences between stayers and leavers. Finally, the mean summative rating GRM is significantly higher (14% of a standard deviation) among stayers at the school level. We also find positive differences between mean GRM among stayers versus leavers at the district and state levels, but they are not statistically significant.

4. Empirical Approach

Our analyses address the question: to what extent are reference ratings predictive of teacher retention at the school, district, and state levels? Below, we outline our approach to answering this question and addressing how the relationship between reference ratings and retention may vary according to rater type (e.g., principal, colleague, university supervisor…),

whether a teacher is novice or experienced, and school level. We also describe our approach to addressing potential bias resulting from selection into the sample.

4.1 Predictive Validity

To assess the relationship between teacher applicant ratings and teacher retention, we estimate models predicting the probability of retention at the school, district, and state levels, with standard errors clustered at the applicant level:

$$f(p_{it}) = \alpha + \beta_1 R_{ijt} + \beta_2 S_{it} + \beta_3 D_{it} + \gamma_t + \varepsilon_{ijt}, \quad (2)$$

where p_{it} is the probability that teacher i in year t is retained in year $t+1$. These models are estimated either as logistic regressions or as linear probability models.

The primary variable of interest, R_{ijt}, is the rating of applicant i by reference j, prior to employment year t, represented as either a categorical variable or the summative ratings measure GRM derived above. We include a vector of controls for the characteristics of each teacher's school, S_{it}, including indicators for school level and the percentage of students eligible for free or reduced-price lunch (FRL).[11] To account for the fact that there are more opportunities for within-district mobility in larger school districts, we control for the number of schools in each teacher's district (D_{it}). In some specifications, we also control for a vector of teacher characteristics T_j (race/ethnicity, experience level indicators, and holding an advanced degree).[12]

[11] We also estimate specifications controlling for the percentage of students who are under-represented minorities and average student achievement on standardized tests (which are strongly correlated with the percentage of students eligible for FRL) and find very similar results.

[12] We do not include these teacher characteristics in our preferred specification because we are interested in learning whether about the extent to which reference ratings send a signal about applicant quality and qualifications such as prior experience and degree level may be incorporated into a references' assessments of applicants.

As noted in prior work examining the relationship between reference ratings and teacher performance (Goldhaber et al., 2023a), differences in ratings standards across references are likely to be a significant source of variation in the distribution of ratings. As such, to account for rater-driven variation, we estimate the model in equation (2) with rater-fixed effects on the subsample of observations where the reference has rated two or more applicants. This subsample is restricted to ratings from 442 of the 2,569 unique raters in the analytic sample.

It is also possible that the extent to which reference ratings are predictive of retention varies according to applicant characteristics. To assess whether reference ratings are differentially predictive for novice and experienced applicants, we estimate equation (2) with an interaction term for whether the applicant is a novice teacher: $GRM_{ij} \times Novice_i$. We similarly use interaction terms to explore whether the relationship between reference ratings and retention differs for applicants hired into school-level types of *elementary*, *secondary*, or *other*.

Finally, to examine the possibility that different types of references are more or less effective at assessing applicants we estimate equation (2) with a rater-type interaction term ($GRM_{ij} \times RaterType_{ij}$), which allows the coefficient β_3 to vary according to rater type. We also include a series of rater-type indicators in these models to examine whether having one or more ratings from a particular type of rater is predictive of retention.

4.2 Addressing Sources of Potential Bias

The models described in Section 4.1 estimate the extent to which reference ratings are predictive of teacher retention. But there are concerns about two potential sources of bias. The first arises from selection into the sample. Specifically, we do not observe retention outcomes of applicants who are not subsequently observed in a classroom teaching position in the WA public

K-12 teacher workforce.[13] Reference ratings are predictive of selection into the sample—the average summative rating GRM of applicants in the analytic sample is 9% of a standard deviation higher than the average rating of excluded applicants. As such, we might anticipate a downward bias in the estimated relationship between reference ratings and retention, assuming that applicants who are hired in spite of low ratings tend to have unobserved attributes that are valued by hiring officials and predictive of retention (e.g., whether the applicant is a good fit).

Following Goldhaber et al. (2020), we assess the potential bias introduced by selection into the sample by implementing a bounding exercise adapted from Lee (2009) by Carrell et al. (2018). As noted above, the relationship between reference ratings and selection into the sample is positive. We calculate the average summative rating $\overline{GRM_i}$ for each applicant and the median value of $\overline{GRM_i}$ across all applicants, GRM_{med}, and estimate the following logit model of selection into the sample:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \overline{GRM_i}\beta_1 + A_i\beta_2 + \gamma_t + \varepsilon_{ijt}, \qquad (3)$$

Where p_i is the probability of being in the analytic sample, A_i is a vector of applicant characteristics, and γ_t is a year fixed effect. We then randomly and incrementally exclude above-average applicants (for whom $\overline{GRM_i} > GRM_{med}$) from the analytic sample until the coefficient $\hat{\beta}_1$ approaches zero (i.e., is less than 0.002). We then estimate the model specified in equation (2), saving the estimated coefficient on GRM_{ij}, and iterate this process 500 times. We use the distribution of point estimates from these 500 iterations to assess the range of potential estimates possible under sample selection that is uncorrelated with the variable of interest.[14]

[13] Additionally, as described above in Section 3.2, we restrict our analytic sample to teachers observed working in a different school than in the prior year and exclude teachers associated with multiple schools.

[14] Following Goldhaber et al. (2023a), we also sought to examine potential bias from selection into the sample by estimating a Heckit model. However, the same instrumental variable, which measured the amount of competition for open positions faced by each applicant, was not statistically significant in the first stage of the model. We found that

A second source of bias we might worry about is associated with the matching of teacher applicants to particular schools. For example, we might worry that stronger applicants (with more positive reference ratings) are more likely to be employed in positions in schools with working conditions that lead to greater retention (Boyd, Grossman, et al., 2011; Burkhauser, 2017; Geiger and Pivovarova, 2018). Our estimates of teacher ratings would be biased upward if working conditions (or other unobserved factors) influencing retention are correlated with ratings and not accounted for by control variables in the model. We assess this potential source of bias by analyzing the correlation between rates of retention *in the hiring school in the year prior to applicants being hired* and the reference ratings of hired applicants. A positive correlation between prior-year retention rates in the hiring schools and reference ratings would suggest that the sorting of applicants into schools is a potential source of bias. We also estimate our primary regression models with a control for prior-year school retention.

5. Results

We present our findings on the relationship between reference ratings and teacher retention, and how that relationship is moderated by applicant type, school level, and rater type in Section 5.1. In Section 5.2 we examine evidence of potential bias from selection into the sample and matching between teachers and schools.

5.1 Reference Ratings and Teacher Retention

As noted in Section 4.1, we estimate the models specified in equation (2) as either logistic regressions or linear probability models. For ease of interpretation, we present results

this lack of significance was driven by two sample restrictions applied in the current paper that were not applied in Goldhaber et al. (2023a): we exclude teachers associated with multiple schools and teachers working in the same school as in the prior year. This results in 98 applicants being classified as "not selected" in the first stage of the Heckit model in spite of their have a status of *Hired* in the SPS job application data.

from the estimation of linear probability models in this section, but estimates obtained from logistic regressions are nearly identical (see Table A2 in the Appendix).

The relationship between reference ratings and retention outcomes is presented in Table 2, with the primary specifications in Panel A and the rater fixed-effect specifications in Panel B.[15] Each model includes controls for school characteristics, the number of schools in the school district, and school-year fixed effects.[16] While we do not report our findings on the relationship between these controls and retention in our regression tables, we find that the school-level and the proportion of students in a school eligible for free or reduced-price lunch are not predictive of retention and that number of schools in a district is predictive of lower rates of school-level retention. Turning to the variable of interest, we find that the summative ratings measure GRM is predictive of retention at the school level: specifically, a one-standard-deviation change in GRM is associated with a 3.2-percentage point increase in retention (column (1)). When in column (2) we substitute the categorical ratings on the overall ratings criterion (see Figure A1) for the *GRM* measure, we also find that ratings are predictive of school retention—receiving a rating of *Among the Best* is associated with a rate of retention 6.2 percentage points higher than the *Very Good* (reference category) and 12.0 percentage points higher than the lowest rating category of *Average/Below Average*.[17] Neither ratings measure is predictive of retention at the district or state level.

[15] As described in Section 4.1, each regression model is estimated with standard errors clustered at the teacher level. When we cluster standard errors at the school level, results are nearly identical.

[16] As noted above, we do not control for teacher characteristics in our preferred specifications. When we introduce controls for teacher experience, race/ethnicity, and holding an advanced degree, we find that effect sizes are attenuated: the coefficient on GRM is 0.023* vs. 0.031**. More experienced teachers (with 2 to 5 or 6+ years of experience) are significantly more likely to be retained at the school level than are novice teachers, but holding an advanced degree and ethnicity are not predictive of retention.

[17] For the purposes of analysis, we merge the bottom two ratings categories on because they are sparse, comprising only 5% of ratings in the analytic sample.

As noted above, we are able to introduce rater fixed effects for the subsample of raters who have rated one or more applicants in the analytic sample. As shown in Panel B of Table 2. We find that the summative ratings measure GRM is significantly predictive of retention at the school, district, and state levels and that the magnitudes of the coefficients on GRM are substantially larger than in our primary specifications. A one-standard deviation increase in GRM is associated with an 8.5 percentage point increase in retention at the school level, and 5.6 and 5.2 percentage point increases in retention at the district and state levels, respectively. We also find large effect sizes for the categorical rating on the *Overall* criterion. Applicants rated in the top two categories have probabilities of retention between 13.6 and 19.6 percentage points higher than applicants rated as *Very Good* and the rater fixed effect sample restriction itself appears to contribute to the larger coefficient estimates but does not explain the majority of the increase.[18] These findings suggest that much of the variation in the ratings of applicants is driven by the raters themselves rather than differences in the quality of applicants and is consistent with prior work that found that the reference ratings exhibited relatively low levels of inter-rater reliability (Goldhaber et al., 2021). It is also consistent with an earlier analysis of the relationship between reference ratings and teacher performance, which found substantially stronger relationships in rater fixed effect specifications (Goldhaber et al., 2023a).

To examine whether there is heterogeneity in the relationship between reference ratings in retention related to teacher or job characteristics, we allow the coefficient on GRM to vary according to whether a teacher is a novice (Panel A of Table 3) and according to school level (Panel B of Table 3). Models including rater fixed effects are presented in columns (2), (4), and (6). Regarding novice versus experienced teachers, we fail to find any significant differences in

[18] When we estimate our primary specifications using the rater fixed effect subsample, we obtain coefficients on GRM of 0.042**, 0.018, and 0.024* from the school, district, and state models respectively.

the relationship between GRM and any level of retention in our full-sample regressions. In the rater fixed effect models, we find that ratings of novice applicants are slightly more predictive of retention at the school level (0.106** vs. 0.073**) and significantly more predictive of retention at the district level (0.093** vs. 0.038).

We assess heterogeneity across school levels in Panel B of Table 3. We find that ratings of elementary teachers are predictive of retention at the school level but fail to find any other statistically significant relationship between ratings and retention in our full-sample specifications (columns (1), (3), and (5)). In our rater fixed effect specifications, we find that ratings of elementary teachers are predictive of retention at the school, district, and state levels, and that ratings of teachers in *Other* school types are predictive of school-level retention and ratings of high school teachers are predictive of retention at the state level.

It is possible that ratings from some types of references are more or less predictive of retention than others. To explore this possibility, we allow the coefficient on GRM to vary according to reference type (results are available in Table A3 in the Appendix).[19] With the exception of the coefficient on $GRM * Principal$ (which is identical in magnitude to the coefficient on GRM in Panel A of Table 2), the GRM-reference type interaction terms are not statistically significantly different from zero, nor from one another. Consistent with the preceding results, we find stronger relationships between these interaction terms and retention under rater fixed effect specifications. To assess whether receiving a rating from a particular type of reference is itself predictive of retention, we also include indicators for whether an applicant

[19] When submitting a reference rating, references are asked to indicate their relationship to the applicant by selecting one of the following options: *Principal/Other Supervisor*, *Instructional Coach/Dept. Chair*, *Cooperating Teacher*, *University Supervisor*, or *Other*.

received one or more ratings from each type of rater. Here too, we fail to find evidence of a significant relationship to retention.

The analyses above have focused on two summative reference ratings measures: GRM (which incorporates information from the six rating criteria) and the categorical ratings on the *Overall* criterion. We also model the relationship between retention and each individual evaluation criterion represented as a categorical variable. We estimate a separate regression model for each criterion because they are strongly correlated with one another and as shown in prior work (Goldhaber et al., 2021), they load onto a single factor. As shown in Figure 1, we find that receiving a top rating of *Among the best (top 1%)* is significantly predictive of school-level retention relative to the reference category of *Very good* for the criteria *Challenges Students* and *Classroom Management* ($p < 0.05$) and marginally predictive for the criteria *Classroom Engagement* and *Instructional Skills* ($p < 0.10$). The criteria *Working with Diverse Groups of Students* and *Interpersonal Skills* are not predictive of school retention. Consistent with our findings for the *Overall* criterion, none of the individual criteria are significantly predictive of district or state-level retention.[20] When we estimate these models with rater fixed effects, however, we find that each individual criterion is predictive of retention at the school level and that the only criterion not predictive of retention at the district and state levels is *Interpersonal Skills* (see Figure A4 to A6 in the Appendix.

Finally, we explore whether the criteria in which applicants are rated as *Strongest* or *Weakest* are significantly predictive of retention. We fail to find evidence that any criterion being identified as an applicant's *Strongest* or *Weakest* competency is significantly more or less predictive of retention vis-à-vis the other competencies (see Table A4 in the Appendix). This is

[20] Estimates from the district and state-level retention models are represented in Figures A2 and A3 in the Appendix and estimates from rater fixed effect specifications are presented in Figures A4 to A6.

consistent with prior work that failed to find any relationship between the *Strongest/Weakest* ratings and teacher performance (Goldhaber et al., 2023b).

5.2 Addressing Threats to Causal Interpretation of Ratings Results

As discussed above, our results may be influenced by two potential sources of bias: selection into the sample and the matching of stronger applicants to high-retention schools. To assess the extent to which our results may be biased by selection into the sample, we conduct the bounding exercise described in Section 4.2. Specifically, we incrementally and randomly exclude applicants with above-average ratings from the regression sample until our variable of interest, GRM, is no longer predictive of selection into the sample. We then estimate our primary regression models predicting retention at the school, district, and state levels. We iterate this process 500 times to estimate the range of coefficient estimates that would occur when selection into the sample is uncorrelated with the variable of interest.

The results from this bounding exercise are presented in Table 4. For reference, we present the baseline results (coefficients on GRM from columns (1), (3), and (5) of Table 2) in column (1). In column (2), we present the means of the coefficients on GRM estimated over the 500 iterations and the 95% confidence intervals of the coefficient estimates. We fail to find evidence that our results suffer from sample selection bias.[21] While the relationship between GRM and selection into the sample is statistically significant, it is worth noting that this relationship appears to be relatively modest. The proportion of applicants that need to be randomly excluded from the sample for the coefficient on GRM in the selection model to approach zero was small, averaging only 13%. One reason for this may be that nearly half of the

[21] The means across iterations of the coefficient estimates are very similar to the point estimates from our primary model specifications.

applicants in the analytic sample are employed in a district other than SPS such that our sample includes both hired and unhired applicants.

To address the concern that stronger applicants may tend to select into schools with healthier work environments (and higher rates of retention) such that the relationship between reference ratings and retention is biased upward, we examine the correlation between GRM and prior-year school retention rates. We find a correlation of 0.025, suggesting that this is unlikely to be a source of bias. Additionally, we estimate our primary regression models in Panel A of Table 2 controlling for prior-year school retention rates and obtain coefficients on GRM that are nearly identical (0.032**, 0.013, and 0.010 for school, district and state-level retention, respectively).

6. Policy Implications and Conclusions

One of the key issues that school officials face when making hiring decisions is making judgments about whether teacher candidates are likely to stay in the positions for which they are hired. In this paper we have explored the potential for a low-cost survey for professional references to add information useful for predicting teacher applicants' propensity of retention. Despite the ubiquity of asking job applicants for references, the predictive validity of information collected from references has received relatively little attention.

Our primary results show that the reference ratings are modestly predictive of retention at the school level (though not at the district or state levels), with a one standard deviation change in our summative ratings measure associated with a 3.2-percentage point increase in retention and receiving a top rating on the *Overall* criterion associated with a 6.2-percentage point increase in retention relative to a rating of *Very Good*. These findings are comparable to those of Bruno and Strunk (2019) and Jacob et al. (2018) who found effect sizes of 1.6 and 4.4 percentage points in examining the relationship between applicant screening scores and school-level retention

(both studies failed to find a significant relationship to district-level retention). It is worth noting, however, that obtaining the applicant screening scores analyzed in these papers was comparatively time-consuming and expensive—for example, in both studies, the applicant screening process involved the scoring of sample lessons by HR professionals.[22] In contrast, as noted in Goldhaber et al. (2023a), we implemented an automated system for collecting ratings of applicants from references at a one-time cost of $2,000.

When we estimate models with rater fixed effects, our findings on the relationship between reference ratings and retention are stronger: we find that a one-standard deviation change in our summative ratings measure is predictive of school-level retention that is 8.5 percentage points higher and district- and state-level retention that is over 5 percentage points higher. These stronger relationships are consistent with previous work on the relationship between reference ratings and teacher performance (Goldhaber et al., 2024a). The potential for reference ratings to meaningfully inform hiring decisions is hampered by low levels of inter-rater reliability; put another way, differences in standards amongst the professional references (i.e., those filling out the ratings) are themselves an important source of error. But this also points to an opportunity. It may be that districts could improve inter-rater reliability by restructuring the ratings instrument to, for instance, provide raters with more guidance about how to think about ways to judge teacher applicants. School district hiring managers might also be able to compare applicant reference ratings to other ratings of applicants performed by the same references. In our analytic sample, this would be possible for about 40% of the ratings and about 60% of applicants.

[22] Jacob et al. (2018) estimated total marginal costs to DCPS to be in the range of $70-200K per year.

The results show that the ratings are predictive of teacher retention and hold up to threats of causal interpretation based on sample selection and non-random matching of teacher applicants to schools. But the findings do not provide much intuition about the value of the information or whether districts could use it to do better in selecting amongst applicants. To help make the predictive power of the models more concrete, we assess their predictive accuracy when different amounts of applicant information are accounted for. Specifically, we generate predicted probabilities of school-level retention based on four different model specifications estimated at the applicant level:[23] 1) a baseline model that only controls for district size, school characteristics, and year fixed effects; 2) baseline controls plus controls for teacher characteristics that would be readily observable to hiring managers—these are, a categorical control for teacher experience and an indicator for holding an advanced degree; 3) baseline controls plus the average summative ratings measure $\overline{GRM}_i$ of each applicant; 4) baseline controls with both the teacher characteristics and $\overline{GRM}_i$. We then assign each teacher a random number between zero and one drawn from a uniform distribution. If the predicted probability that an individual teacher stays exceeds the random number, then that teacher is classified as a stayer. We calculate the percentage of predictions that are correct under each model and iterate the process 500 times.[24] The accuracy rates of the predictions average 59.29% under the baseline model, 59.85% when teacher characteristics are controlled for, 59.60% when reference ratings are included, and 60.01%. when both teacher characteristics and reference ratings are included.

[23] We estimate the regression models at the applicant level (rather than the applicant-rating level) and because it allows us to generate predictions that are constant within applicant. As such, we use each applicant's average reference rating.

[24] This methodology is akin to the "count" calculation of a pseudo R^2, where the continuous predicted probabilities are also transformed into binary (0,1) variables, which are then compared to the actual (0,1) outcomes. The primary difference is that the count pseudo R^2 treats any record with a predicted probability of 0.5 or greater as having a predicted outcome of 1 and any record with a predicted probability less than 0.5 as having a 0 (see: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/), whereas we rely on the random number from the uniform distribution.

Relative to the baseline model, the teacher characteristics add more to the predictive accuracy than do the reference ratings, but the addition of reference ratings do provide an incremental improvement in accuracy relative to accounting for teacher characteristics alone.

The above increases in predictive accuracy are modest, but it is worth considering some of the context that makes them valuable. In particular, the predictive accuracy simulations have a limitation that may downplay the utility of the ratings: they do not account for more nuanced information that is available to hiring officials such as who the rater is and how the ratings relate to other available applicant information such as letters of recommendation and details about applicant work history. And the ratings are easy for hiring officials to interpret. Hiring officials can quickly assess straightforward numbers about teacher applicants, rather than trying to weigh different applicant attributes (as the regression does in predicting attrition). Finally, it is worth emphasizing that this is a very low-cost intervention and even small increases in teacher retention are likely to be cost-effective given the significant financial recruitment and onboarding costs associated with hiring new teachers (Barnes et al., 2007).[25]

The benefit of collecting ratings of teacher applicants also depends on the scope for changing which applicants are hired. SPS has roughly 5 applicants for each available teaching slot, suggesting a degree of choice when it comes to considering who to hire.[26] Given this, we next consider the retention implications of SPS making different choices about who to hire.

Figure 2 shows the distribution of average *GRM* ratings for applicants newly hired into SPS (excluding internal transfers) and those who were not hired by SPS (as we show in Table 1,

[25] Importantly, the direct financial implications of reducing turnover likely understate the benefits of greater retention as various studies have found that the churn of teachers is disruptive to the educational process and tends to negatively impact student achievement independent of any changes in teacher quality (Atteberry et al., 2017; Ronfeldt et al., 2014).

[26] We identify 3,206 unique applicants in the reference ratings data who we are able to link to certification data. We observe 598 of these applicants employed in a new school in SPS during the study period giving us a ratio of 5.3 to 1.

many of these applicants are employed in other districts).[27] While the distribution of average *GRM* among hired applicants (the dashed blue line) is to the right of the *GRM* distribution for non-hired applicants (the solid red line), there is a considerable amount of overlap in these distributions. This implies that SPS could have hired a large number of applicants with average ratings that are higher than many of the newly hired applicants.[28] For example, about 36% of non-hired applicants have an average *GRM* rating that exceeded the average rating of those that were hired (the averages are the dashed drop lines from the kernel densities).

Considering the implications of counterfactual hiring scenarios is an inherently speculative exercise. We would expect applicants hired despite relatively low ratings to have unobserved (to us) attributes that would be correlated with classroom success (including a high propensity of retention), and vice versa for non-hired applicants. But, as a counterfactual scenario, suppose that some of the highly rated candidates were hired in place of some of the hired applicants who received relatively low ratings. Specifically, if the *hired* teacher applicants with below the mean value of *GRM* ratings (the blue dashed line) in SPS were replaced by a randomly selected group of *non-hired* teacher applicants who had *GRM* ratings above the mean of those hired,[29] the average GRM for the full group of hired-into-SPS teachers would be 0.68 higher. Based on the estimated coefficient on *GRM* from Table 2 (column (1)), this suggests that school-level retention of teachers hired into SPS would increase by $0.68 * 3.2 = 2.2$ percentage points.

[27] The implicit assumption is that the applicants not observed as hired were not offered jobs that they refused. This is a reasonable assumption since in prior work, we found that over 95% of applicants offered a position accepted it or another position in SPS that year (Goldhaber et al., 2017).
[28] SPS's job application data show that only 3% of job offers are declined.
[29] Note that this is relatively conservative. In particular, this involves a swap of 283 teacher applicants, but there were 1,205 teacher applicants not hired who were above the mean. The average GRM for the randomly selected group of non-hired teacher applicants is 0.66, whereas the average GRM below-average hired applicants being replace is -0.47.

The findings we describe support the idea of surveying professional references to obtain assessments of teacher applicants. However, the lower interrater reliability of the ratings also suggests room for improvement. Moreover, we do not know whether the reference ratings provide information beyond what districts could derive from other sources of information in an application package, such as recommendation letters, or whether the ratings of the professional references impact teacher hiring decisions. These are both issues for future research.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2023). When Should You Adjust Standard Errors for Clustering? *The Quarterly Journal of Economics*, *138*(1), 1–35. https://doi.org/10.1093/qje/qjac038

Atteberry, A., Loeb, S., & Wyckoff, J. (2017). Teacher Churning: Reassignment Rates and Implications for Student Achievement. *Educational Evaluation and Policy Analysis*, *39*(1), 3–30. https://doi.org/10.3102/0162373716659929

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The Cost of Teacher Turnover in Five School Districts: A Pilot Study*.

Borman, G. D., & Dowling, N. M. (2008). Teacher Attrition and Retention: A Meta-Analytic and Narrative Review of the Research. *Review of Educational Research*, *78*(3), 367–409. https://doi.org/10.3102/0034654308321455

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The Influence of School Administrators on Teacher Retention Decisions. *American Educational Research Journal*, *48*(2), 303–333. https://doi.org/10.3102/0002831210380788

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The role of teacher quality in retention and hiring: Using applications to transfer to uncover preferences of teachers and schools. *Journal of Policy Analysis and Management*, *30*(1), 88–110. https://doi.org/10.1002/pam

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 Conditions? *Educational Evaluation and Policy Analysis*, *39*(1), 126–145. https://doi.org/10.3102/0162373716668028

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

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

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

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

Geiger, T., & Pivovarova, M. (2018). The effects of working conditions on teacher retention. *Teachers and Teaching*, *24*(6), 604–625. https://doi.org/10.1080/13540602.2018.1457524

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. (2023a). 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., Kasman, M., Quince, V., Theobald, R., & Wolff, M. (2023b). How did it get this way? Disentangling the sources of teacher quality gaps through agent-based modeling. *Social Science Research*, *116*(November). https://doi.org/https://doi.org/10.1016/j.ssresearch.2023.102941

Gray, L., & Taie, S. (2015). *Public School Teacher Attrition and Mobility in the First Five Years: Results From the First Through Fifth Waves of the 2007–08 Beginning Teacher Longitudinal Study*. https://files.eric.ed.gov/fulltext/ED556348.pdf

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The Market for Teacher Quality* (No. 11145; NBER Working Paper Series, Issue No. 11154). National Bureau of Economic Research. http://www.nber.org/papers/w11154.pdf

Harris, D. N., Rutledge, S. A., Ingle, W. K., & Thompson, C. C. (2010). Mix and Match : What Principals Really Look for When Hiring Teachers. *Education Finance and Policy*, *5*(2), 228–246.

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

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Vol. 398). John Wiley & Sons.

Ingersoll, R. M., & Strong, M. (2011). The Impact of Induction and Mentoring Programs for Beginning Teachers. *Review of Educational Research*, *81*(2), 201–233. https://doi.org/10.3102/0034654311403323

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

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

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

Nodoye, A., Ritzhaupt, A. D., & Parker, M. A. (2012). Use of ePortfolios in K-12 Teacher Hiring in North Carolina: Perspectives of School Principals. *International Journal of Education Policy & Leadership*, *7*(4), 1–10.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

Rockoff, J. E. (2004). The Impact of Individual Teachers on Students' Achievement: Evidence from Panel Data. *American Economic Review*, *94*(2), 247–252. http://www.ingentaconnect.com/content/aea/aer/2004/00000094/00000002/art00046

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

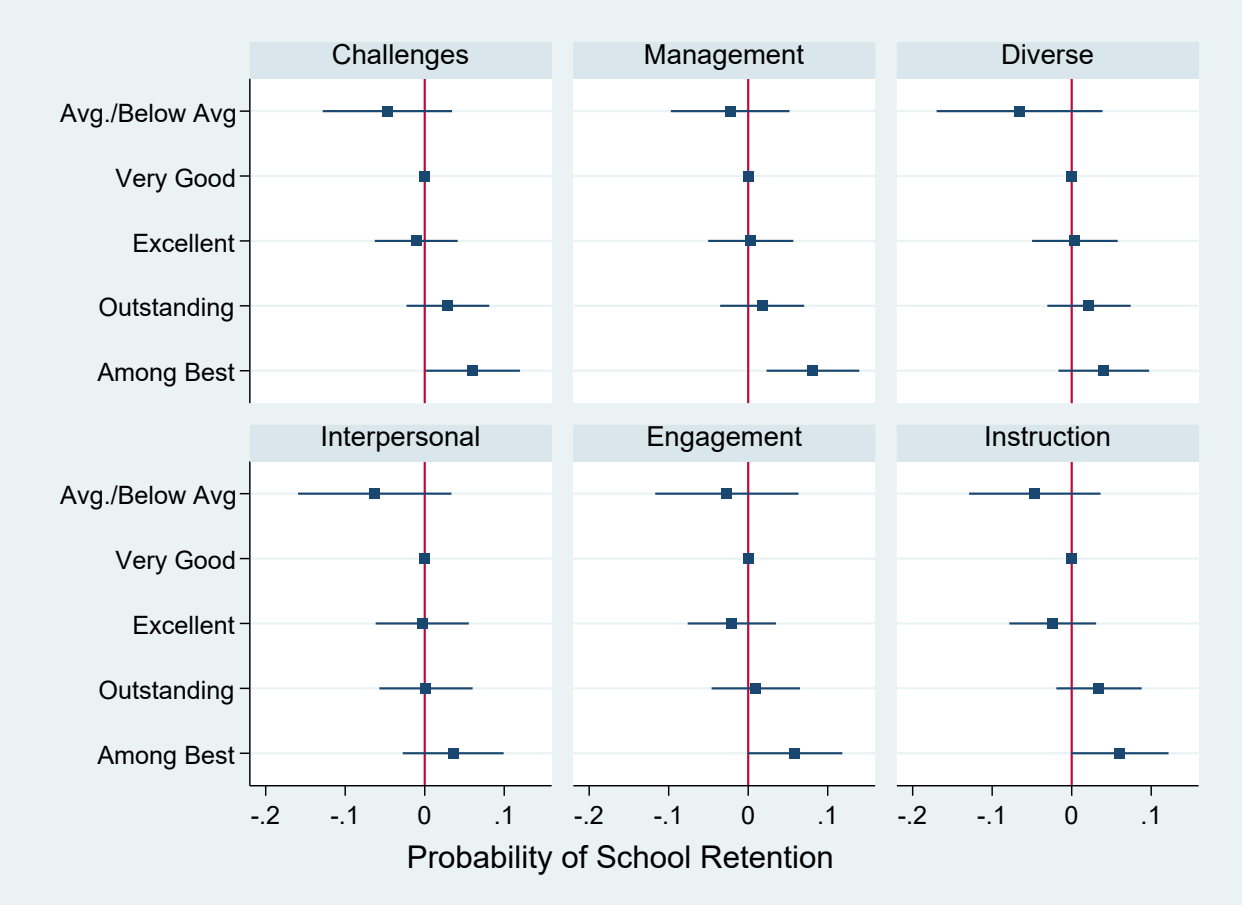
Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225. https://doi.org/https://doi.org/10.1037/apl0000405

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

Sorensen, L. C., & Ladd, H. F. (2020). The Hidden Costs of Teacher Turnover. *AERA Open*, *6*(1), 1–24. https://doi.org/10.1177/2332858420905812

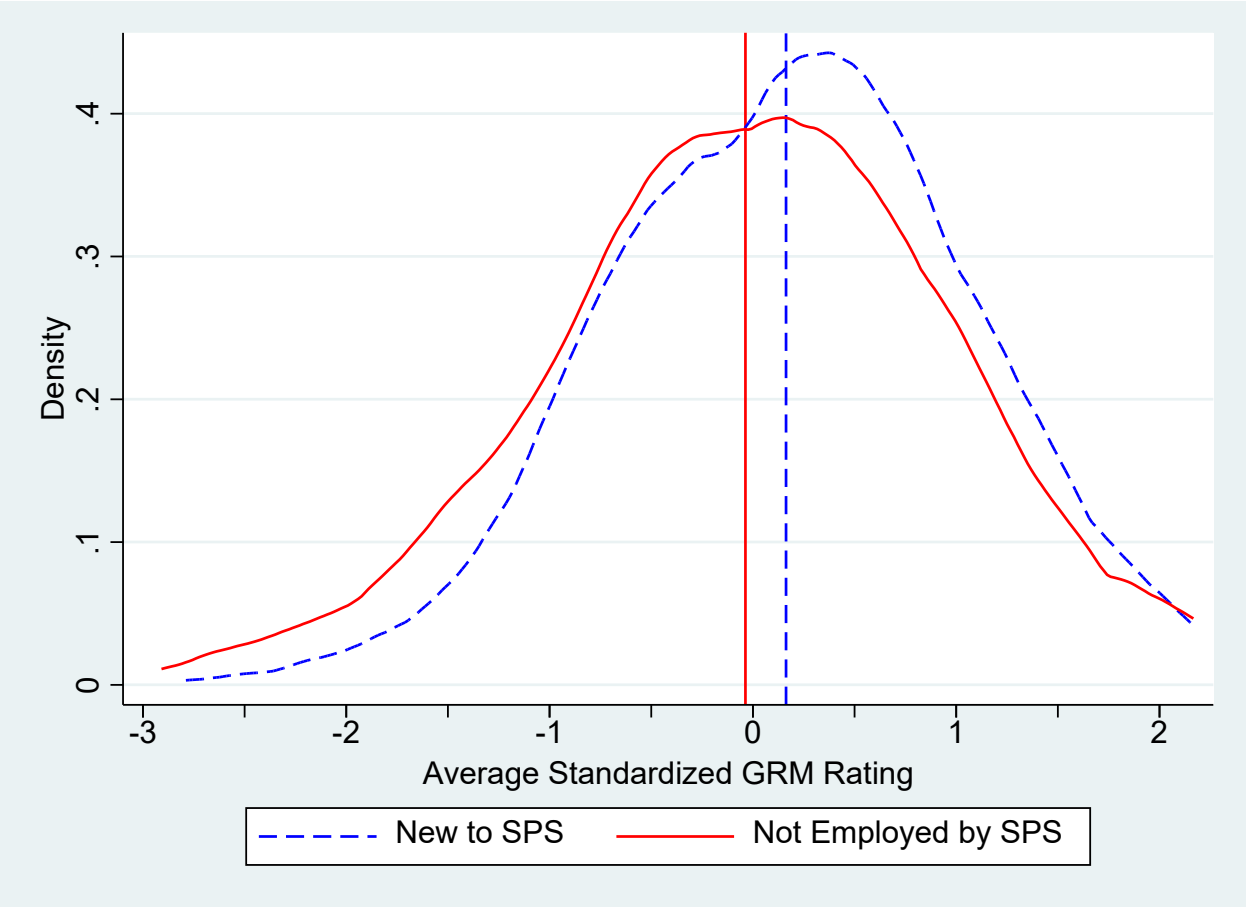
Figures and Tables

Figure 1. Coefficients on Individual Criteria – School Retention



Notes: Each plot represents a separate regression model, estimated as a linear probability model. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 2. Distribution of Average GRM Rating



Note. The average summative ratings measure GRM is calculated for each applicant and standardized (0, 1). The plot is restricted to observations of applicants new to SPS and applicants not employed by SPS, excluding applicants already employed by SPS and internal transfers within SPS.

Table 1. Analytic Sample Descriptive Statistics

		Difference of: $Mean(stayers) - Mean(leavers)$		
	All	School	District	State
	(1)	(2)	(3)	(4)
Teacher Characteristics				
Experience: 0 years	0.32	**-0.06**	-0.02	0.02
Experience: 1 year	0.16	-0.01	0.02	0.02
Experience: 2 to 5 years	0.23	-0.01	-0.03	-0.03
Experience: 6 plus years	0.28	**0.08**	0.03	-0.02
Female	0.75	-0.03	-0.01	0.03
White	0.97	-0.01	-0.01	0.00
Advanced Degree	0.47	-0.02	-0.04	-0.05
School Characteristics				
In SPS	0.53	**-0.09**	0.03	0.00
Level: elementary	0.56	**-0.07**	0.01	-0.01
Level: middle	0.17	0.00	-0.02	-0.04
Level: high	0.18	**0.05**	0.03	0.04
Level: other	0.09	0.01	-0.02	0.01
Percent FRL	56.52	**-3.80**	-1.30	0.67
Percent URM	0.31	**-0.02**	-0.02	0.01
Math score (avg.)	-0.14	**0.11**	0.09	0.05
ELA score (avg.)	-0.20	0.08	0.05	-0.02
Observations	1,124	1,124	1,124	1,124
Reference Ratings				
Overall: Below Avg/Avg	0.05	**-0.02**	**-0.02**	**-0.04**
Overall: Very Good	0.14	-0.01	-0.02	-0.02
Overall: Excellent	0.23	-0.02	0.01	0.04
Overall: Outstanding	0.35	0.02	0.01	0.01
Overall: Among Best	0.23	**0.04**	0.01	0.01
GRM Ratings Measure	0.00	**0.14**	0.08	0.07
Observations	3,532	3,532	3,532	3,532

Notes: The analytic sample is restricted to applicants with reference ratings subsequently observed in a classroom teaching position in a new or different school than in the previous school year and for whom we observe the full set of control variables used in our regression analyses. The reference ratings are restricted to those for which a response is given on each evaluation criterion (excluding ratings with one or more responses of *No basis for judgment)*. Statistics for teacher and school characteristics are calculated at the ratings level and statistics for reference ratings at the ratings level. Individuals who apply for positions in multiple hiring years are treated as distinct applicants. Percent FRL refers to the percentage of students eligible to receive free or reduced lunch. ELA refers to English language arts. *GRM* refers to the summative reference ratings measure derived from the estimation of a graded response model described in Section 3.1. Columns 2 to 4 report the difference of the mean characteristics of stayers versus leavers. Bold text indicates that difference in mean values between stayers and leavers is statistically significant at the 5% level according to a two-group t-test

Table 2. Predicting Retention

	School		District		WA K-12 Public Teacher Workforce	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Primary Specifications						
GRM	0.032**		0.012		0.009	
	(0.010)		(0.009)		(0.008)	
Overall Criterion						
Avg./Below Avg.		-0.058		-0.039		-0.066
		(0.044)		(0.042)		(0.039)
Very Good		Ref.		Ref.		Ref.
		-		-		-
Excellent (top 10%)		0.009		0.031		0.036
		(0.027)		(0.024)		(0.020)
Outstanding (top 5%)		0.032		0.020		0.019
		(0.027)		(0.025)		(0.021)
Among best (top 1%)		0.062*		0.027		0.021
		(0.031)		(0.027)		(0.024)
Observations	3,532	3,532	3,532	3,532	3,532	3,532
Clusters/Teachers	1,124	1,124	1,124	1,124	1,124	1,124
R^2	0.083	0.082	0.014	0.015	0.012	0.016
Panel B: Rater Fixed-Effect Specifications						
GRM	0.085***		0.056**		0.052**	
	(0.023)		(0.019)		(0.017)	
Overall Criterion						
Avg./Below Avg.		0.012		0.079		0.028
		(0.075)		(0.063)		(0.056)
Very Good		Ref.		Ref.		Ref.
		-		-		-
Excellent (top 10%)		0.034		0.083		0.088*
		(0.054)		(0.050)		(0.040)
Outstanding (top 5%)		0.196***		0.159**		0.136**
		(0.059)		(0.053)		(0.044)
Among best (top 1%)		0.159*		0.156*		0.168**
		(0.071)		(0.064)		(0.053)
Observations	1,412	1,412	1,412	1,412	1,412	1,412
Clusters/Teachers	677	677	677	677	677	677
R^2	0.372	0.373	0.381	0.391	0.357	0.361

Notes: GRM is the standardized summative reference ratings measure described in Section 2.3. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3. Heterogeneity in Predicting Retention by Experience Level and School Level

	School		District		WA K-12 Public Teacher Workforce	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: GRM by Novice/Experienced						
GRM*Novice	0.033	0.106**	0.023	0.093**	0.011	0.050*
	(0.020)	(0.038)	(0.019)	(0.030)	(0.016)	(0.022)
GRM*Experienced	0.031**	0.073**	0.009	0.038	0.008	0.053*
	(0.011)	(0.027)	(0.010)	(0.023)	(0.010)	(0.021)
Rater Fixed Effects		X		X		X
Observations	3,532	1,412	3,532	1,412	3,532	1,412
Clusters (unique apps.)	1,124	677	1,124	677	1,124	677
R^2	0.083	0.373	0.015	0.383	0.012	0.357
Panel B: GRM by School Level						
GRM*Elementary	0.041**	0.105***	0.013	0.061**	0.013	0.062**
	(0.013)	(0.029)	(0.012)	(0.024)	(0.011)	(0.022)
GRM*Middle	0.007	0.011	0.005	0.020	0.006	0.008
	(0.025)	(0.039)	(0.023)	(0.034)	(0.021)	(0.031)
GRM*High	0.010	0.032	0.009	0.049	0.004	0.076*
	(0.021)	(0.049)	(0.019)	(0.047)	(0.016)	(0.035)
GRM*Other	0.055	0.186**	0.025	0.120	-0.001	0.035
	(0.034)	(0.064)	(0.031)	(0.066)	(0.021)	(0.044)
Rater Fixed Effects		X		X		X
Observations	3,532	1,412	3,532	1,412	3,532	1,412
Clusters (unique apps)	1,124	677	1,124	677	1,124	677
R^2	0.084	0.377	0.015	0.383	0.013	0.361

Notes: *GRM* is the standardized summative reference ratings measure described in Section 2.3. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4. Carrell et al. Bounds for Sample Selection Bias

	Baseline	Bounding exercise	
	Pr(Retention)	Average	95% Confidence Interval
Coefficients on GRM	(1)	(2)	(3)
School retention	0.032**	0.031	[0.0249, 0.0374]
	(0.010)		
District retention	0.012	0.0133	[0.0076, 0.0196]
	(0.009)		
State retention	0.009	0.0108	[0.0058, 0.0164]
	(0.008)		
Random draws		500	
School controls	Yes	Yes	
Teacher controls	No	No	

Notes: The baseline estimates presented in column (1) are from the regression models presented in Panel A of Table 2. The estimates presented in column (2) are means of the coefficients on GRM estimated over the 500 iterations on samples with randomly excluded applicants. The number of observations in the bounding exercise regression models averaged 3,070 (87% of the full analytic sample). Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The 95% confidence intervals are the 2.5th and 97.5th percentile values of the coefficients estimated over 500 iterations.

Appendix. Supplemental Figures and Tables

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 ▼

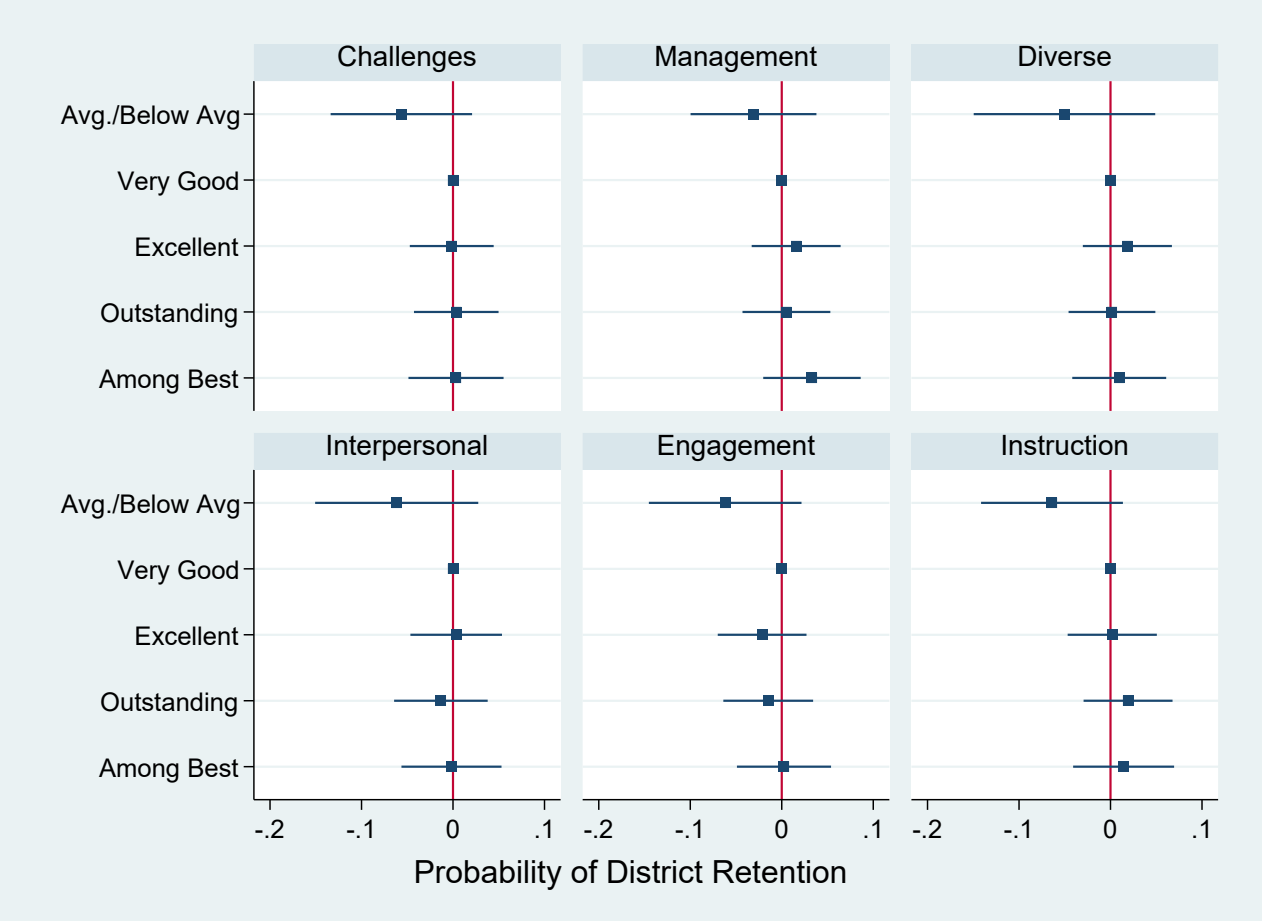
Overall, how would you rate the candidate?

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)

Submit

Figure A2. Coefficients on Individual Criteria – District Retention

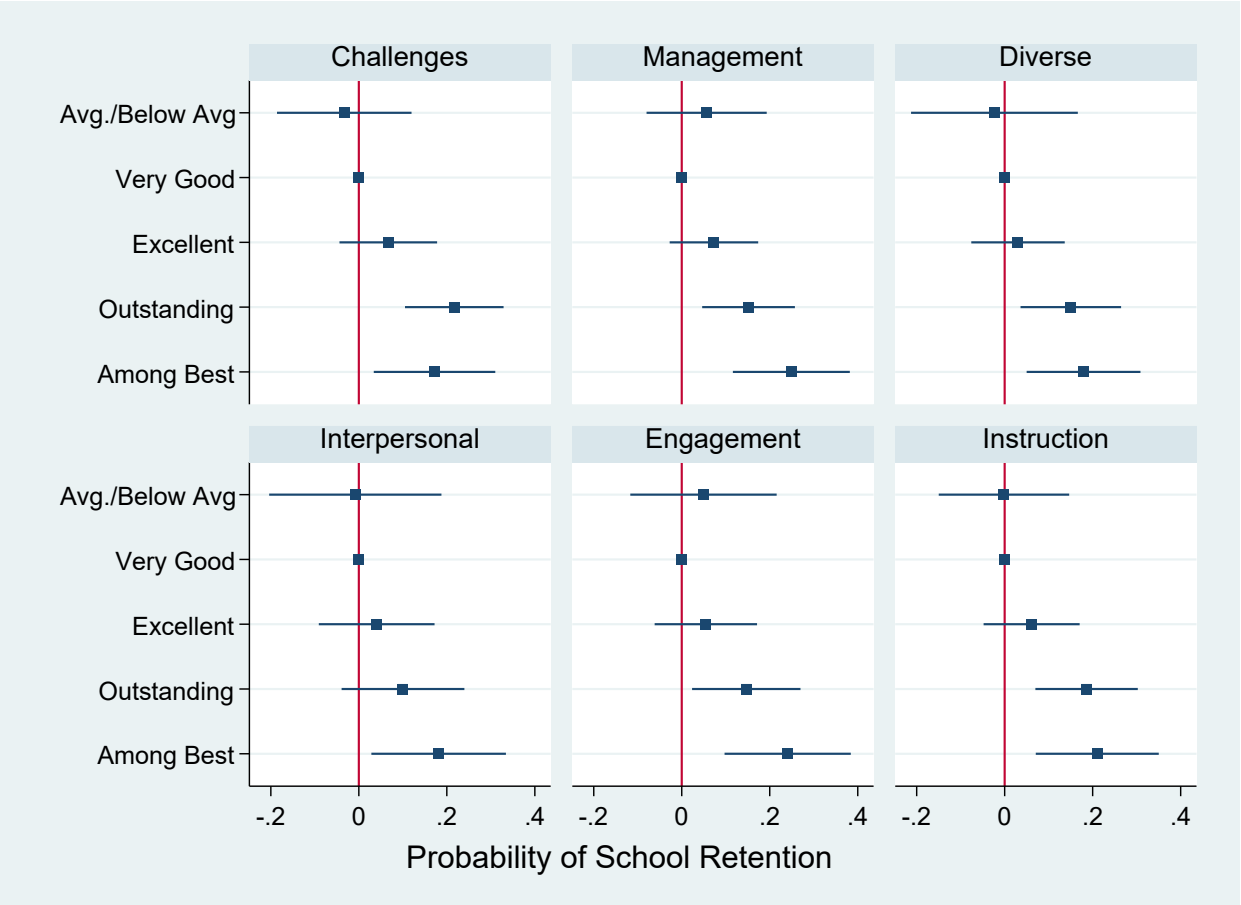


Notes: Each plot represents a separate regression model, estimated as a linear probability model. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A3. Coefficients on Individual Criteria – State Retention

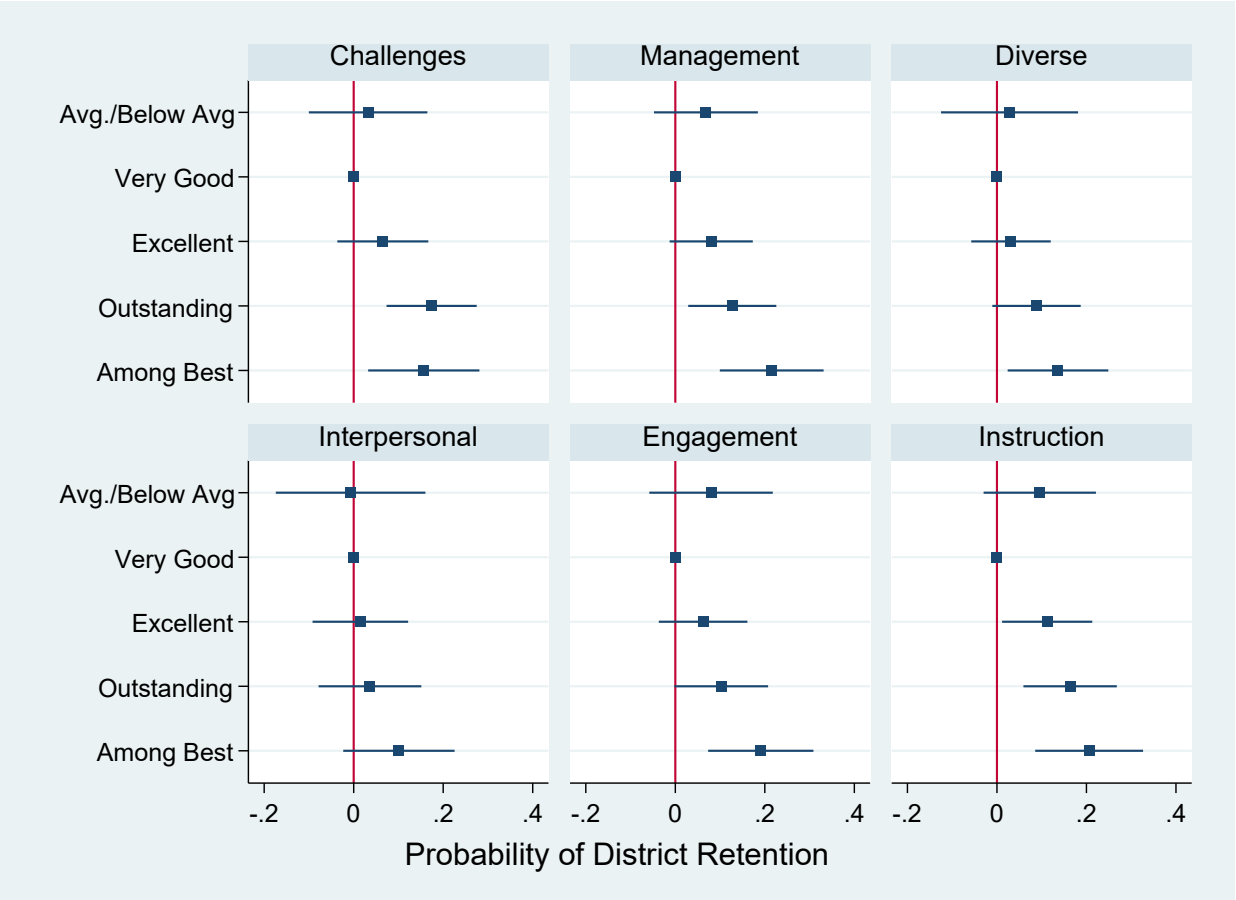
Notes: Each plot represents a separate regression model, estimated as a linear probability model. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A4. Coefficients on Individual Criteria – School Retention with Rater Fixed Effects



Notes: Each plot represents a separate regression model, estimated as a linear probability model estimated with rater fixed effects. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A5. Coefficients on Individual Criteria – District Retention with Rater Fixed Effects



Notes: Each plot represents a separate regression model, estimated as a linear probability model estimated with rater fixed effects. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure A6. Coefficients on Individual Criteria – State Retention with Rater Fixed Effects

Notes: Each plot represents a separate regression model, estimated as a linear probability model estimated with rater fixed effects. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

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. Predicting Retention using Logistic Regression

	School		District		WA K-12 Public Teacher Workforce	
	(1)	(2)	(3)	(4)	(5)	(6)
GRM	0.155**		0.076		0.078	
	(0.050)		(0.056)		(0.073)	
Overall Criterion						
Avg./Below Avg.		-0.270		-0.270		-0.218
		(0.205)		(0.205)		(0.223)
Very Good		Ref.		Ref.		Ref.
		-		-		-
Excellent (top 10%)		0.045		0.186		0.329
		(0.132)		(0.143)		(0.175)
Outstanding (top 5%)		0.156		0.118		0.160
		(0.134)		(0.148)		(0.183)
Among best (top 1%)		0.317*		0.163		0.180
		(0.155)		(0.164)		(0.203)
Observations	3,532	3,532	3,532	3,532	3,532	3,532
Clusters/Teachers	1,124	1,124	1,124	1,124	1,124	1,124
Pseudo-R^2	0.066	0.066	0.014	0.015	0.016	0.021

Notes: GRM is the standardized summative reference ratings measure described in Section 2.3. Coefficients are presented as log-odds. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3. Heterogeneity in Predicting Retention by Reference Type

	School		District		WA K-12 Public Teacher Workforce	
	(1)	(2)	(3)	(4)	(5)	(6)
Reference Type						
Principal	0.037	0.080	0.062	0.117*	0.037	0.037
	(0.036)	(0.059)	(0.033)	(0.048)	(0.025)	(0.038)
Instr. Coach/Dept. Chair	0.060	0.127	0.039	0.087	0.043	0.062
	(0.039)	(0.069)	(0.036)	(0.053)	(0.027)	(0.037)
Colleague	0.011	0.057	0.049	0.091*	0.033	0.033
	(0.032)	(0.053)	(0.028)	(0.044)	(0.023)	(0.030)
Cooperating Teacher	-0.029	-0.026	0.058	0.041	0.056*	0.060
	(0.039)	(0.063)	(0.032)	(0.047)	(0.028)	(0.032)
University Supervisor	-0.043	-0.051	-0.034	-0.044	0.004	0.008
	(0.041)	(0.068)	(0.036)	(0.051)	(0.031)	(0.035)
Other	0.032	0.035	0.030	0.095*	0.045	0.098***
	(0.037)	(0.060)	(0.032)	(0.046)	(0.023)	(0.029)
GRM by Reference Type						
GRM*Principal	0.032*	0.071*	0.022	0.044	0.020	0.048*
	(0.015)	(0.029)	(0.014)	(0.024)	(0.013)	(0.022)
GRM*Instr. Coach/Dept. Chair	0.019	0.003	0.009	0.024	0.011	0.104
	(0.025)	(0.067)	(0.022)	(0.066)	(0.013)	(0.057)
GRM*Colleague	0.023	0.091	0.000	0.028	-0.003	0.041
	(0.017)	(0.059)	(0.015)	(0.044)	(0.013)	(0.036)
GRM*Cooperating Teacher	0.021	0.148*	0.020	0.117*	0.016	0.091*
	(0.021)	(0.064)	(0.019)	(0.047)	(0.016)	(0.035)
GRM*University Supervisor	0.044	0.054	0.021	0.060	0.015	0.054
	(0.025)	(0.046)	(0.023)	(0.041)	(0.018)	(0.030)
GRM*Other	0.022	0.213*	-0.012	0.147*	-0.001	0.110
	(0.028)	(0.086)	(0.024)	(0.073)	(0.018)	(0.074)
Rater Fixed Effects		X		X		X
Observations	3,532	1,412	3,532	1,412	3,532	1,412
Clusters (unique applicants)	1,124	677	1,124	677	1,124	677
R-Squared	0.094	0.386	0.025	0.404	0.024	0.377

Notes: GRM is the standardized summative reference ratings measure described in Section 2.3. Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The rater fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A4. Predicting Retention Using Strongest/Weakest Ratings

	School Retention		District Retention		State Retention	
	Strongest	Weakest	Strongest	Weakest	Strongest	Weakest
	(1)	(2)	(3)	(4)	(5)	(6)
Challenges Students	0.033	-0.012	0.012	-0.008	0.011	0.007
	(0.035)	(0.029)	(0.032)	(0.026)	(0.027)	(0.022)
Classroom Management	0.005	0.007	0.014	-0.006	0.026	0.030
	(0.030)	(0.030)	(0.028)	(0.027)	(0.024)	(0.023)
Instructional Skills	0.004	0.029	0.004	0.019	0.004	0.020
	(0.026)	(0.035)	(0.023)	(0.032)	(0.020)	(0.028)
Interpersonal Skills	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
	-	-	-	-	-	-
Student Engagement	-0.016	0.034	-0.003	0.003	0.003	-0.003
	(0.026)	(0.037)	(0.023)	(0.032)	(0.020)	(0.028)
Working w/ Diverse Grps.	-0.006	-0.002	-0.002	-0.017	-0.004	-0.005
	(0.028)	(0.029)	(0.025)	(0.027)	(0.023)	(0.024)
Observations	3,532	3,532	3,532	3,532	3,532	3,532
Clusters (unique applicants)	1,124	1,124	1,124	1,124	1,124	1,124
R-Squared	0.079	0.079	0.014	0.014	0.013	0.013

Notes: Each regression includes a categorical control for school level, the percentage of students at the teacher's school who are eligible for free or reduced-price lunch, the number of schools in the teacher's school district, and a school year fixed effect. The fixed-effects sample is restricted to ratings from references who submitted two or more ratings of an applicant during the study period. Standard errors are clustered at the teacher level. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$