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

TRACKING EVERY STUDENT'S LEARNING EVERY YEAR

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

AIR

Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools

DAN GOLDHABER
CYRUS GROUT
NICK HUNTINGTON-KLEIN

WORKING PAPER 120 • DECEMBER 2014

Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools

Dan Goldhaber
American Institutes for Research
University of Washington

Cyrus Grout
Center for Education Data & Research

Nick Huntington-Klein
University of Washington

Contents

Acknowledgments

We acknowledge support from the Institute of Education Science's Researcher–Practitioner Grant Program (Grant #R305C130030) and from the National Center for Analysis of Longitudinal Data in Education Research (CALDER), funded through grant #R305A060018 to the American Institutes for Research from the Institutes of Education Sciences, U.S. Department of Education. This research has benefitted from the helpful input of Heather Hill, Angela Jones, Susanna Loeb, Jonah Rockoff, Mary Templeton, and members of Spokane Public Schools' stakeholder group, as well as from data management and analysis support by Andrew Katz, Malcolm Wolff, Patricia Martinkova, Roddy Theobald, and Nate Brown. We also thank all data providers, including Spokane Public Schools, the Washington Office of the Superintendent of Public Instruction, and the Washington State Information Processing Collective. Any and all errors are solely the responsibility of the study's authors, and the views expressed are those of the authors and should not be attributed to their institutions, the study's funders, or the agencies supplying data.

CALDER • American Institutes for Research
1000 Thomas Jefferson Street NW, Washington, DC 20007
202-403-5796 • www.caldercenter.org

Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools

Dan Goldhaber, Cyrus Grout, and Nick Huntington-Klein

November 2014

Abstract

Evidence suggests that teacher hiring in public schools is ad hoc and often fails to result in good selection among applicants. Some districts use structured selection instruments in the hiring process, but we know little about the efficacy of such tools. In this paper, we evaluate the ability of applicant selection tools used by the Spokane Public Schools to predict three outcomes: measures of teachers' value-added contributions to student learning, teacher absence behavior, and attrition rates. We observe all applicants to the district and are therefore able to estimate sample selection-corrected models, using random tally errors in selection instruments and differences in the quality of competition across job postings. These two factors influence the probability of being hired by Spokane Public Schools but are unrelated to measures of teacher performance. We find that the screening instruments predict teacher value added in student achievement and teacher attrition but not teacher absences. A one-standard-deviation increase in screening scores is associated with an increase of between 0.03 and 0.07 standard deviations in student achievement and a decrease in teacher attrition of 2.5 percentage points.

1. Introduction

Teachers can have a profound effect on student outcomes. Empirical estimates find that a one-standard-deviation increase in teacher effectiveness raises student test achievement by 0.10 to 0.25 standard deviations and that teachers can affect long-term student outcomes, such as college-going behavior and labor market earnings. Not surprisingly, the last decade has seen a considerable amount of research and policy attention directed toward interventions that can improve the quality of the teacher workforce. These interventions include efforts to increase quality through alternative certification, new processes of evaluation and feedback, professional development, provision of performance incentives, and more recently, focus on preservice teacher training.

There is far less research or policy focus on the choices school systems make in the teacher hiring process. This is surprising for several reasons. First, there is a large and growing body of economic research pointing to the importance of the hiring process, particularly for sectors of the economy that rely heavily on human capital. Second, many school districts have a significant amount of choice among job candidates, but once they have hired teachers (particularly if tenured), removing ineffective ones can be quite costly (National Council on Teacher Quality, 2014; Treu, 2014). Third, the credentials that are generally used to determine employment eligibility and reward in-service teachers tend to be only weakly correlated with teacher effectiveness, meaning that required state employment screens and in-service financial rewards are unlikely to lead to productive labor market sorting. When a teacher is hired, districts are making what may turn out to be a large, long-term financial commitment; it is sensible to make sure that the recruitment and selection process works well.

In this paper, we analyze the relationship between two teacher selection rubrics that are used during the teacher hiring process in Spokane Public Schools (SPS) and three teacher outcomes: value-added measures of effectiveness, teacher absence behavior, and the likelihood of attrition. All three of

these measures are arguably quite important. Value-added measures of teacher effectiveness have been found to be predictive of students' future test achievement and long-term outcomes. Evidence also suggests that teacher absences are negatively related to students' test achievement (Clotfelter, Ladd, & Vigdor, 2009; Herrmann & Rockoff, 2011; Miller, Murnane, & Willet, 2008) and may, in addition, have broader impact on students and schools (Clotfelter et al.). Finally, teacher attrition has important implications for both district administrative costs and student achievement.

Three aspects of our study make it unique. Unlike previous studies of the hiring process, ours observes employment outcomes for applicants who are hired by SPS and applicants who are not hired by SPS but are then employed in other public school districts in Washington State. The observation of the teachers not hired by SPS allows for a more comprehensive analysis than would be possible if we were limited to observations of teachers who perform well enough to progress through the entire hiring process. In fact, as we detail below, the ability to observe teachers who end up employed outside of SPS affects the interpretation of the value of the screening rubrics. The reason is that the relationship between applicant ratings on the rubrics and teacher outcomes varies along the applicant performance distribution and, not surprisingly, Spokane tends to employ teachers who score toward the top of the distribution.

Second, we observe whether a prospective teacher has been offered a job (which he or she may reject), not just whether a teacher is employed in a position. Thus, we are able to distinguish between job applicant nonmatches (i.e., an applicant is not employed in SPS) that result from employee preferences and those that result from employer preferences. We find that 95% of prospective Spokane teachers who receive an offer from Spokane accept the offer. The difference between a teacher who ends up employed in Spokane and a teacher who does not is, then, largely a decision on the district's part.

Finally, we are able to correct for selection bias that may arise from not being able to observe the outcomes of teachers who are not employed in public schools in Washington. Our selection-corrected estimates exploit the fact that a nontrivial proportion of the summative scores teachers receive on the selection instruments are incorrectly computed because of procedural oversight or arithmetic mistakes, as well as the differing amounts of competition faced by applicants when applying for SPS jobs; these factors are assumed to influence the likelihood of being hired but should not otherwise be related to teacher quality.

We find that teachers hired by Spokane are more effective (as measured by value added) than applicants who end up employed by a different school district in Washington. Hired applicants also tend to be absent more often and are less likely to leave their district. The summative ratings of the two selection instruments used by Spokane are associated with these differences. Screening scores have strong relationships with both teacher value added and teacher attrition, and the magnitudes of these relationships are educationally meaningful: A one-standard-deviation increase in screening scores is associated with an increase of about 0.07 standard deviations of student math achievement, a marginally significant increase of 0.03 to 0.05 standard deviations of student reading achievement, and a decrease in teacher attrition of 2.5 percentage points. Correcting for selection for a SPS job does not significantly change the findings, thereby suggesting that teachers who do not end up employed in Washington's public schools do not significantly bias the estimates.

These findings are evidence that public schools can improve the quality of the teacher workforce through the use of well-designed applicant selection tools. However, our analyses of the subcomponents of the instruments show much stronger relationships for some components than for others, implying that the teacher outcomes we assess could be further improved by weighting certain subcomponents, such as Classroom Management, more strongly than others.

The paper proceeds as follows. Section 2 provides background on hiring practices and employee performance, and describes the hiring process in SPS. Section 3 describes the data we used for this study, and provides a descriptive picture of which applicants move through the hiring process in SPS. In Section 4, we describe the econometric methodology, and in Section 5, we describe the results of the analyses. Section 6 discusses the policy implications of our findings and offers some conclusions.

2. Background

2.1 Hiring Practices and Employee Performance

The potential for improving workforce quality through effective hiring practices is broadly supported by research from the field of personnel economics (Heneman & Judge, 2003; Shaw & Lazear, 2007) and industrial psychology (see Society for Industrial and Organizational Psychology, 2014, for an overview). Studies analyzing the validity of screening and selection processes, using a wide variety of employers and employee groups, generally find that screening tools based on biographical data (experience and training) improve the process of worker selection. A meta-analysis by McDaniel, Schmidt, and Hunter (1988) of assessments of education and experience finds that different types of screening scores have average correlations with measures of job performance between 0.11 and 0.45. Another meta-analysis by Bliesener (1996) calculates adjusted average correlations between 0.15 and 0.32. While suggestive, these meta-analyses are hardly definitive, since the they draw from are all based on relationships between information about employees who were hired and job performance. These studies do not observe a nonhired counterfactual and the reported correlations do not account for sample selection, relying generally on corrections for restricted range.

Perhaps contrary to prevailing wisdom, public schools often have a significant amount of choice among potential teachers.[1] How do schools select from among applicants to fill teaching positions? The Schools and Staffing Survey, administered by the National Center for Education Statistics, suggests that schools rely on teacher licensure and graduation from a state-approved teacher education institution as important screens (U.S. Department of Education, 1997), even though there is little evidence that these credentials are a good proxy for teacher quality (Goldhaber & Brewer, 2000; Glazerman, Mayer, & Decker, 2006). Regarding the interview process, Harris, Rutledge, Ingle, and Thompson (2010) extensively interview school principals and find that they desire a mix of teacher characteristics, including experience, enthusiasm, pedagogical skills, and content knowledge, as well as "organizational fit" in terms of achieving a school-level balance of experience, race, and gender. Indicators of academic proficiency appear to be given little weight in hiring decisions. Using school principal survey data, Mason and Schroeder (2010) find that principals rate references (both written and verbal) and first impressions as more important sources of applicant information than a candidate's portfolio of qualifications. Ballou (1996) and Hinrichs (2013) find that school hiring decisions are not very sensitive to measures of an applicant's academic proficiency, such as college selectivity and SAT scores, whereas Boyd, Lankford, Loeb, and Wycoff (2013) find that these decisions are sensitive to proficiency.

Under the current approach to hiring in public schools, there is mixed evidence about whether the best applicants are the ones hired. A few studies look at whether schools identify the best applicants in terms of a direct measure of teacher effectiveness and value added, as opposed to the data available to the school system at the time of hiring, but here too, the evidence is mixed. Hanushek, Kain, O'Brien, and Rivkin (2005) analyze whether schools that offer higher levels of compensation (in the form of both

[1] Ingersoll and Perda (2010) report that, in 2000 (when the teacher labor market was relatively tight), the ratio of teachers in the supply pipeline to the number of teachers leaving through retirement and attrition was more than 2 to 1. Similarly, Strauss, Bowes, Marks, and Plesco (2000) find evidence of excess supply in Pennsylvania, where 75% of districts hiring for various subject areas had at least 3 applicants per position. In elementary education, mathematics, English, and social studies, there were at least 10 applicants per position.

salary and work environment) tend to hire more effective teachers. They find that these schools (generally suburban schools) are more likely to hire teachers with advanced degrees but do not find a relationship between hiring and value-added measures of effectiveness.

Staiger and Rockoff (2010) found evidence consistent with Hanushek et al., using a natural experiment that occurred in the Los Angeles Unified School District (LAUSD). In 1997, California provided a financial incentive to limit kindergarten through third grade (K–3) class sizes to fewer than 20 children. As a consequence, LAUSD more than doubled its annual hiring of new elementary school teachers during the following 5 years. During this time, teacher pay did not increase and the proportion of new hires without teaching credentials increased from 59% to 72%. Value-added estimates of elementary teacher effectiveness showed no evidence of a decrease, suggesting that, prior to the change, the district had not been selecting teachers from the top end of the distribution. Boyd, Lankford, Loeb, Ronfeldt, and Wyckoff (2011), on the other hand, examine within-district transfer applicants and find that schools tend to hire teachers with higher value added.[2]

If there is room to improve the hiring process, then it is possible that screening tools or other information available during hiring can help. Screening tools are used by at least some (unknown number of) school systems.[3] The Haberman Star Teacher Pre-Screener and the Gallup TeacherInsight Assessment, for instance, are commercially available instruments that purport to be able to predict teachers' future success through assessments of applicants' values and beliefs. But few studies of these instruments are independent of their developers or those with commercial interests (Metzger & Wu,

[2] More recent evidence (Cannata et al., 2014) suggests that some school districts have begun to incorporate direct measures of teacher effectiveness data in the hiring process, but to the best of our knowledge, there is no assessment of such practices.

[3] Ebmeier and Ng (2006) offer a review of interview-based screening practices in the context of teaching and find that standard interview practices (e.g., an interview with a principal) do not relate strongly to teacher performance, but interview-based *tools* that help interviewers reach unbiased judgments about applicant attitudes and teaching practices can help schools make good decisions about which applicants to hire.

2008; Rockoff, Jacob, Kane, & Staiger, 2011), and there is relatively little evidence of their efficacy based on independent analysis.

To the best of our knowledge, only two studies directly connect the information available to employers about teacher applicants to the productivity of those applicants once they are employed, and both potentially suffer from selection bias in that they are limited to assessments of those applicants who are hired. Dobbie (2011) investigates the link between information used to select Teach for America (TFA) members and the future impact that TFA teachers have on student achievement (in the first year of teaching).[4] He finds that a one-standard-deviation change in an index that averages the standardized measures used to select TFA applicants is predicted to increase student achievement by about 0.15 standard deviations on a math exam (although the findings for reading achievement were smaller and not statistically significant).

Rockoff et al. (2011) examine the extent to which traditional (e.g., degree and major, passage of license exam) and nontraditional information about teacher applicants (e.g., measures of conscientiousness, extraversion, personal efficacy) are related to teacher value-added scores, subjective teacher ratings, teacher absences, and retention. The authors find that few *individual* metrics are significant predictors of teacher effectiveness but that a one standard deviation increase in distilled measures of cognitive and noncognitive skills derived using factor analysis is associated with significant increases in student math achievement (0.033 standard deviations).[5] The variation of predicted value added, using both traditional and nontraditional information, explains about 12% of the expected variance in teacher effectiveness, compared with about 4%, using only traditional information. The

[4] For more research on TFA corps members' success in the classroom compared with the success of those who enter teaching through other routes, see, for instance, Boyd, Grossman, Lankford, Loeb, and Wyckoff (2006); Glazerman et al. (2006); and Xu, Hannaway, and Taylor (2011).

[5] Also, cognitive skills are found to be predictive of retention, and noncognitive skills are significantly predictive of mentor evaluation scores. Neither factor is significantly predictive of teacher absences, although both exhibit a negative relationship. It is important to note that the commercial instrument that is used—the Haberman Star Teacher Pre-Screener—predicts teacher performance on its own, with a one standard deviation in screening scores associated with a marginally significant 0.023 standard deviation increase in math achievement.

Rockoff et al. findings suggest that the quality of the workforce could be improved by collecting, and using, additional information about prospective teachers when they apply for teaching positions.

In sum, the relatively thin literature on the teacher hiring process generally supports the notion that it is possible to improve the quality of the workforce through better hiring. In particular, the literature is *suggestive* that one way to do this is by using screening instruments.

2.2 The Hiring Process in Spokane Public Schools

During the 2008–2009 through 2012–2013 hiring years, SPS received applications from 2,669 applicants for 521 positions, which were filled via their standard hiring process.[6] Below, we outline the hiring process to give context to the data-generating process and describe the set of applicant information available to SPS hiring officials.

The hiring process in SPS, outlined in **Figure 1**, includes four stages following a school job posting:

1. Intake of applications
2. Twenty-one-point prescreening of potential applicants by Human Resources (HR) hiring officials
3. Sixty-point screening of applicants by school-level hiring officials
4. In-person interview and hiring decision

Job applications are submitted through an online applicant management system.[7] Applicants submit information on their education, qualifications (i.e., certifications and endorsements), experience (teaching, student teaching, and other professional positions), letters of recommendation (a minimum of three), and narrative statements. When a job posting is closed, an applicant list is created and all applications are checked for completeness.

Job applicants must progress through two stages of screening before they become eligible to interview for a position. The first stage, "prescreening," conducted under the direction of HR, which a 21-point scoring rubric with three subcomponents: Experience, Depth of Skills, and Quality of

[6] Spokane Public Schools (SPS) is the largest school district in eastern Washington and the second largest in the state. In 2012, the district included 34 elementary schools, 6 middle schools, and 5 high schools, and employed approximately 1,800 teachers, who instructed 28,800 students.
[7] Applicants can use one application profile to apply for multiple jobs.

Recommendations (hereafter referred to as Recommendations).[8] An applicant's prescreening score is not directly associated with his or her application for any particular job in the district. Typically, an applicant is only screened the first time he or she applies for an SPS position, or when new qualifications have been obtained (e.g., an applicant receives a new teaching endorsement). The applicant is scored on each criterion on a scale of 1 to 6, and the Recommendations score is then multiplied by 1.5. (**Table 1** describes what the screener looks for in an applicant's profile to score each criterion.) Scores of 1–2, 3–4, and 5–6 indicate the finding of, respectively, "some," "satisfactory," and "strong" evidence that the criterion is an area of strength for the applicant.[9]

The primary use of the 21-point score is to narrow the applicant pool to a manageable size before the second stage of screening, which is conducted at the school level and led by school principals. In the second stage, principals request that HR provide a list of applicants for consideration. These requests typically specify a cutoff score on the 21-point screening. For example, a principal hiring a second-grade teacher could request a list of all applicants with an endorsement in Elementary Education and a screening score of greater than 17. Principals and HR followed these procedures closely. We observe very few cases of what might be seen as noncompliance, such as the advancement of applicants who do not achieve the requisite point cutoff. We will discuss the issue of potential noncompliance in more depth in Section 3.

The third stage of the hiring process used to select the candidates who will receive in-person interviews. This job-level screen uses a 60-point rubric with 10 evaluation criteria and the same 1 to 6

[8] Starting in spring 2013, SPS replaced the 21-point screening score with the Marzano Teacher Evaluation Model, which is used in many districts across the country and is one of the evaluation methods recommended by the Washington State Teacher Evaluation Project (see http://www.tpep-wa.org/). The data used for the analyses presented in this paper do not overlap with the implementation of the Marzano-based rubric.

[9] Applicants are not screened unless they have fully completed an application. Also, internal applicants are not generally prescreened when applying for a different job within the school district. School principals are notified of each internal application but are only *required* to consider the two most senior qualified applicants, who are automatically granted interviews.

scoring scale as in the 21-point rubric (see an example of the 60-point screening form in **Appendix A**).[10] The screening criteria are outlined in **Table 2**. Each applicant may be evaluated by one or more screeners, and principals use the screening scores as the basis for selecting the applicants to interview.

The final stage of hiring consists of in-person interviews. The principal assembles a team to interview the applicants selected through earlier stages of the screening process. The principal has discretion over the content of the interview in terms of what questions are asked, how the interview is structured, and how many people are on the hiring team (teams typically have four members). In contrast to the prior parts of the screening process, in the third stage there are no set interview evaluation criteria established centrally by SPS's HR Department. Following the completion of the interviews, the principal submits a "request to hire" form, along with copies of the interview questions and scoring sheets. After background checks and with final approval from the district's HR Department, a job offer can be made.

3. Data and Descriptive Portrait of Applicants Moving Through the SPS Hiring Process

We link administrative data at the applicant, job, teacher, student, and school levels. These data allow us to analyze the relationship between teacher screening scores and the three teacher measures – teacher value added, absences, and attrition – controlling for school, teacher, and student characteristics.

Student data come from the statewide Core Student Record System (CSRS). The CSRS data set includes information on student demographics and test scores. Students' teacher assignments are also observed in CSRS and can be used to link students to teachers and schools.

[10] The 10th criterion, Letters of Recommendation, was only added in 2011 but was used only for some jobs in the 2011 or the 2012 hiring years. Below, we discuss how we deal with these inconsistencies across years and jobs.

Teacher and applicant data come from multiple sources. Information on each applicant is provided by SPS, including records of which jobs each applicant applied for, data on applicant characteristics (as described in Section 2.2), scoring of applicants on screening instruments, progression of applicants through the hiring process (whether the applicant is screened, interviewed, offered a job, and ultimately hired), and information on absences of teachers who are hired by Spokane.

We link data on Spokane applicants to statewide teacher data sets, using unique teacher certification numbers. These data include teacher licensure test scores and areas of endorsement, collected by the Professional Education and Standards Board, and teacher absence data for teachers who do not work in Spokane, collected by the Washington School Information Processing Cooperative. We also link applicants to the S-275 personnel report, which provides a record of all certificated employees of public school districts in Washington State, including demographic information, experience level, contract information, and building assignment information. Data on school characteristics come from Public School Universe data generated by the National Center for Education Statistics (NCES).

We study the pool of applicants for all certificated classroom teaching jobs for which Spokane hired teachers using the process outlined in the previous section during the 2009 through 2012 hiring years. A unique "job" refers to an open position available at a given school for a given assignment at a given time in the hiring cycle. Thus, for example, one middle school math teacher opening for which a particular school started hiring in June is a different job from a middle school math teacher opening at the same school in November. A job is a "certificated classroom teaching" job if it requires personnel who hold a valid teaching credential and entails the teacher's spending the majority of his or her time instructing students. In total, Spokane filled 521 job postings fitting this description. There were 2,669 applicants for these jobs. However, many applicants applied for multiple jobs; the average number of applications per

job was 135 for elementary positions and 34 for other positions (see **Table B1** in **Appendix B),** even though the ratio of applicants to jobs was 5.1 overall.

Summary statistics for the screening data are presented in Table 3. We perform most analyses at the applicant-year level, and so the following presentation of data focuses on that observation level. Many of the 2,669 unique individuals applied to SPS in one or more years, generating a total of 4,217 unique applicant-year combinations between 2009 and 2012. However, some of these applicants were not given a 21-point screening score; thus, there were a total of 3,944 unique applicant-year combinations in which applicants were screened at the 21-point stage. If an applicant is given multiple 21-point prescreenings within 1 year, his or her average score within that year is used in the Table 3 calculations. It is important to note that there is a substantial amount of variation in the 21-point and 60-point screening scores, a necessary condition for the instruments to be able to differentiate among applicants of differing quality.

The scores of the individual components that make up the 21-point screening score are available for a subset of 2,672 applicant-year combinations. The second (60-point) screening stage was reached for 1,711 applicant-year observations. The average applicant reaching the 60-point screening stage did so at 3.4 different jobs. The average number of times an applicant was screened per job is 1.2, and an application to a particular job is screened by multiple people approximately 15% of the time.

In some cases, there are important differences between the total "rater score" and the score we can calculate by examining the scoring on the sub-components of the screening instruments, referred to as the "calculated score." These differences can arise when a screener fails to enter a component score (resulting in a missing observation) or makes arithmetic errors (resulting in different totals). The adjusted rater and calculated totals differ on the 21-point score for approximately 19% of the job applications observed in our analyses, and differ on the 60-point score for approximately 8% of the screenings. As we describe below in more detail (Section 4.2), we use these inconsistencies on the 21-

point score as a variable that predicts the probability of an applicant's being hired but that can be excluded from the teacher performance model.

There are also differences in the number of observations for the different criteria. A screener may not enter a score for a particular applicant if there is insufficient information with which to rate the applicant or may forget to fill out a score. For instance, the pattern of missing values for the Preferred Qualifications field suggests that screeners often failed to notice the criterion because it was often on the back side of the screening form.

The missing values for some criteria in the unadjusted scores pose a problem for our analyses because, even though inconsistencies across jobs in the use of the 60-point rubric does not affect which applicants are hired (as long as the inconsistencies do not arise within job screenings), we are comparing screening scores and outcomes across jobs and schools. Therefore, as we show in Table 3, we adjust the scores for cross-school comparability and use adjusted scores in the analyses in Section 4.

The following procedure is used to adjust the blanks that appear for some subcomponents of the 60-point rubric screening scores, so that the applicants' total ratings are comparable across jobs. First, we attempt to identify blanks that result from the rater correctly scoring the component but not writing down the answer: If the rater total is higher than the calculated total by an amount that clearly allows the missing value to be identified, then the missing value is filled in with the difference between the rater and calculated totals. Second, in cases where we cannot infer whether a rater simply failed to record an answer, we correct the rating in one of two ways. In cases where all applicants for a particular job received a zero on the criterion, we replace blank scores with the overall sample mean. In cases where not all applicants for a particular job received a zero on a specific component, we replace the blank with a score of zero. The assumption driving these two adjustments is that, if a criterion is blank for all candidates, it is being systematically excluded as part of the evaluation (whether accidentally or intentionally), and replacing the value with the sample mean simply makes scores more comparable

across jobs. On the other hand, if a criterion is blank for only some applicants and not others, it is being used as part of the overall evaluation, in which case the implicit score for the candidate with a blank is zero.

The above adjustments increase the average rater and calculated totals by about 3.5 points. Most of the individual criteria are not significantly affected (because they had few missing values), but the number of observations for the Preferred Qualifications and Letters of Recommendation criteria increase dramatically. As mentioned above, the Preferred Qualifications field had many blanks because it was often unnoticed on the back of the form, and the Letters of Recommendation field had many blanks because it was introduced midway through the sample period and was not always used after its introduction. In practice, however, it makes little difference to our findings if we use unadjusted scores in analysis. This is not terribly surprising, given the fact that we adjusted less than 1% of the subcomponents on the 21-point rating and less than 10% of the subcomponents on the 60-point rating, most of which were Letters of Recommendation scores.

Table 4 presents the pair-wise correlations of the unadjusted screening scores and their subcomponents. It is interesting to note that the total 21-point and 60-point screening scores are not highly correlated (0.17). However, we do see that the correlations of the most similar categories across the instruments exhibit the highest correlations, although not as high as one might expect (0.19–0.28). Even though both the 21- and 60-point scores are derived from effectively the same information, there are at least two possible reasons for these low correlations. First, the 21-point screening rubric is not job-specific, whereas the 60-point screening rubric is. Second, what the screener is instructed to look for is not precisely the same for both screenings (see Tables 1 and 2). Looking within the 60-point criteria, Classroom Management, Flexibility, Instructional Skills, and Interpersonal Skills are highly coordinated with one another (0.67–0.76). With the exception of Certificate and Education, each 60-point criterion is highly correlated with the total score (0.65–0.75).

Three measures are used to evaluate teacher outcomes: Grades 3 through 8 student performance on Washington State's annual assessments of student learning for math and reading, teacher absences for each day of the week in 2012 and 2013, and teacher retention. Teacher outcome data are linked to the most recent screening scores. So, for example, consider a teacher employed continuously in Washington State from 2009 to 2012. If that teacher applied to Spokane in both 2009 and 2010, then the 2011 and 2012 teacher performance outcomes are linked to the 2010 application, the 2010 outcomes are linked to the 2009 application, and the 2009 outcomes are not used. In this manner, we are able to match 274 applicant-year observations to student test score data and 502 applicant-year observations to absence data. Teacher retention in the district is determined by matching applicants to the S-275 personnel records for the school years ending between 2010 and 2013. For the 2009 applicants, we are able to observe whether a teacher returns after as many as 4 years of service, and for the 2012 applicants we can identify who returns after up to 1 year of service. We are able to match 736 applicant-year observations to certificated employment records in the S-275 data.

Descriptive statistics of applicant data and teacher outcomes over each stage of the hiring process are presented in Table 5. Of the 4,217 applicant-year combinations, 3,944 (or 93%) are prescreened by HR using the 21-point rubric, 1,709 of these (41%) are passed along to schools for consideration where they are scored on the basis of the 60-point rubric, 1,238 are interviewed (29% of the total applicant-years), and 538 (13%) are hired or offered a new job in Spokane. Nearly all (95%) of those offered a position accepted it or another in Spokane that year, suggesting little need to distinguish between offers and hires. An additional 498 (12%) are identified as being employed in a certificated teaching position in a different district in Washington State by October of the same year they applied to Spokane, whether by obtaining a new job or staying in a currently held position. In total, 32% of applications lead to a certificated classroom teaching position in Washington by October of the next school year.

The distribution of applicant characteristics suggests that SPS values familiarity in determining which applicants to advance through the hiring process. Eleven percent of applications each year come from those identified as being employed in a certificated teaching position in Spokane during the application year, but these applicants make up 43% of those who are eventually hired. This is not terribly surprising, in part because those applying for an internal transfer receive preferential treatment in the hiring process because of an agreement between the district and the teacher's union and also because teachers applying for an internal transfer have average Washington Educator Skills Test Basic (WEST-B) scores 0.14 standard deviations higher than others. Similarly, those with student teaching experience in the district are overrepresented among those applicants who are hired or offered a job: 36% of applicants obtained student teaching experience within the district, and these student teachers represent 47% of those who are hired. In total, about 71% of hired teachers had some previous Spokane experience, as an employee, as a student teacher, or both. The distribution of college attended is fairly stable across the stages of the hiring process.

The summary statistics in Table 5 suggest that Spokane's hiring process is effective at selecting high-quality teachers. Average value-added scores generally increase as the application pool narrows, although less dramatically for student reading scores. Average annual and Monday or Friday absences are fairly stable across the stages of the hiring process, and hired applicants average a slightly greater number of absences. The proportion of teachers observed attritting within 1 or 3 years is quite stable through the hiring pipeline, except that those who are hired tend to attrit less often. SPS applicants perform slightly below the state average on the state's licensure exam (WEST-B), but average scores are generally higher among applicants who progress further through the hiring process.

The data described above are generated by the real-world hiring processes of a large public school district. Not surprisingly, the process outlined in Section 2.2 is not always followed to the letter. Mistakes and exceptions can occur at each stage of the process, and we do see some evidence of what

might be a small amount of noncompliance. For example, 0.7% of applications advanced to the 60-point screening stage despite having a 21-point score below the principal's requested cutoff, and 5.4% of applications with 21-point scores above the cutoff were not advanced to the 60-point screening stage and were in the same pool as a lower-scoring applicant who was advanced. Many of these discrepancies can be explained by the presence of formally stated requirements in 90% of principal requests that would allow low-scoring applicants with other useful job-specific qualifications to advance, and would block applicants with acceptable screening scores.

We also find a small number of cases (less than 4% of job applications screened at the school level) in which the lowest 60-point score of an applicant receiving an interview request is lower than the highest score of an applicant not receiving an interview request. This pattern is driven by cases in which an applicant pool is screened and interviewed but no applicant is hired; this results in a second round of screenings.

4. Methods

Our analysis investigates the extent to which the SPS screening instruments are predictive of student achievement, teacher absences, and teacher retention. We describe analytic models for these primary outcomes below and follow with a discussion of several supplemental models.

4.1 Primary Outcome Models

Student achievement

To assess the relationship between teacher scores on the screening instruments and student achievement, we estimate a two-step model. In the first step, we estimate a student achievement model, from which we draw teacher value added. Then we estimate teacher value-added as a function of screening scores. We use a two-step process, rather than including the screening score in the student

achievement model in a one-step process, so that the coefficient on the screening score is not affected by the relationship between screening scores and student assignments[11]:

$$Y_{ijsgt} = \alpha_j + \alpha_1 Y_{i(g-1)(t-1)} + \alpha_2 X_{igt} + \varepsilon_{ijst}^{\alpha} \quad (1)$$

$$\hat{\alpha}_j = \alpha_0' + \alpha_1' SCREEN_{j(tprior)} + \alpha_g' + \alpha_t' + \alpha_y' + \varepsilon_{jst}^{\alpha'} \quad (2)$$

In Equation 1, Y_{ijsgt} is the test score for each student *i* in class with teacher *j* in subject *s* (math or reading), grade *g,* and year *t,* normalized within grade, year, and subject; $Y_{i(g-1)(t-1)}$ is a vector of the student's scores in the previous grade and year in both math and reading, also normalized within grade, year, and subject; X_{igt} is a vector of student attributes in grade *g* and year *t* (gender, race, eligibility for free or reduced-price lunch, English language learner status, gifted status, special education status, learning disability status, migrant status, and homeless status). Equation 1 allows us to calculate teacher value added, represented by the teacher fixed effect α_j.[12]

Teacher value-added estimates $\hat{\alpha}_j$ are the dependent variable in Equation 2. Since $\hat{\alpha}_j$ are estimated values, each observation in Equation 2 is weighted by the inverse of the standard deviation of $\hat{\alpha}_j$.[13] In the second step, we include indicators for grade α_g' and year α_t', and indicators α_y' for the number of years (1, 2, or 3) between the hiring year $tprior$ and the year t, in which performance data are observed, or the "gap."[14] Standard errors are clustered at the teacher level. The coefficient of interest in Equation 2 is α_1', which represents the expected change in teacher effectiveness (and thus the average standardized exam score of their students) associated with a one-standard-deviation increase in that

[11] The primary model results, described below, are qualitatively similar using a two-step or a one-step process. However, results differ when dividing the sample by teachers who worked in Spokane against those who worked elsewhere, as in the selection model. This suggests that the relationship between teacher quality and their students' assigned characteristics differs in and out of the district.

[12] We experiment with a model that includes teacher experience as an additional control, but we report results from a model that does not include teacher experience because experience is one of the categories that is used to determine an applicant's screening score.

[13] Value-added estimates in Table 5 are, instead of being weighted by standard deviations, as in the regression models, "shrunk" using empirical Bayes methods, which is similar to weighting by inverse standard deviation but is more applicable outside a regression context. For more on the use of Empirical Bayes estimates in teacher value-added models see (Aaronson, Barrow, & Sander, 2007; Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008).

[14] The observed difference between t and $tprior$ is typically 1.

teacher's screening score. $SCREEN_{j(tprior)}$ is the screening score for teacher *j* (which varies between specifications) at time *tprior*, which indicates the timing of the most recent screening score, as outlined in the previous section. $SCREEN_{j(tprior)}$ is standardized over the sample.

We estimate six specifications of the model in Equation 1 for math and reading. In Specification 1, $SCREEN_{j(tprior)}$ is the teacher's score on the 21-point prescreening. In Specification 2, $SCREEN_{j(tprior)}$ is one of the components of the 21-point prescreening instrument. In Specification 3, $SCREEN_{j(tprior)}$ is the teacher's adjusted rater total on the 60-point screening. Specification 4 is analogous to Specification 2 but includes each component of the adjusted 60-point screening score in $SCREEN_{j(tprior)}$.[15] In Specification 5, $SCREEN_{j(tprior)}$ includes the teacher's total scores on both the adjusted rater total for the 21-point and 60-point screening scores.

Specification 6 uses derived linear combinations of the 21- and 60-point component scores. These are based on a factor analysis on all 21- and 60-point component scores with a promax rotation to derive two underlying factors of the component scores.[16] In Specification 6, $SCREEN_{j(tprior)}$ consists of these two underlying factors, generated using the calculated loadings from the factor analysis.[17] The estimated factors are presented in **Appendix Table B2** and effectively load completely onto the two different screening scores so that Factor 1 is largely a reweighting of the 60-point score and Factor 2 is a reweighting of the 21-point score.

Teacher absences

[15] We estimate Specifications 2 and 4 separately for each subcomponent, to avoid issues of collinearity between components.

[16] Two factors are used because this is the number of factors with an eigenvalue above 1. These are factors that account for more variance than does a single variable. The promax rotation emphasizes factor weights on particular components and makes the identification of underlying factors more straightforward and understandable.

[17] To avoid issues of sensitivity and to present a simpler version of the factor analysis, we also estimated a version of Specification 6 in which the linearly constructed factor variables were replaced by the average of the components found in each factor to have a loading of more than 0.3. Results were similar.

To assess the relationship between teacher scores on the screening instruments and teacher absences, we estimate the following teacher-level model, using ordinary least squares (OLS) regressions with year and gap indicators and standard errors clustered at the teacher–hiring year level:

$$A_{jkdt} = \beta_0 + \beta_1 T_j + \beta_2 S_{kt} + \beta_3 SCREEN_{j(tprior)} + \beta_t + \beta_y + \varepsilon_{ijst}^{\beta} \quad (3)$$

When multiple years of absence data are available, we use all available years, and if an individual applies in multiple years, we only match his or her most recent application to the absence data (which are available for the 2012 and 2013 school years), as outlined above. In the model in Equation 3, A_{jkdt} is the number of absences for teacher *j* in school *k*, district *d*, and year *t* (the definition of absence varies across specifications); T_j is a vector of teacher characteristics (gender and race); S_{kt} is a vector of school characteristics in year *t* (size, student demographics, level, and location); and $SCREEN_{j(tprior)}$ is the screening score for teacher *j* (which varies across specifications, as described for Equation 2) in year *tprior*. The coefficient of interest in Equation 3, β_3, represents the expected change in teacher absences associated with a one-standard-deviation increase in the teacher's screening score, all else being equal. Standard errors are clustered at the teacher–hiring year level.

We consider two types of teacher absences as the dependent variable in Equation 3. In one set of models, A_{jkdt} is the number of sick days taken by the teacher during year *t*, while in the other set of models, A_{jkdt} is the number of Monday or Friday absences for the teacher during year *t*.[18]

Teacher retention

To assess the relationship between teacher scores on the screening instruments and teacher retention, we estimate logits predicting that teachers leave the district.[19] Specifically, let $p_{jkdt}(y)$ be the probability that teacher *j* in school *k*, district *d*, and year *t* leaves the district y years after being hired. These outcomes are modeled as follows:

[18] There is strong evidence of discretionary absence behavior (e.g., Miller, Murnane, & Willet, 2008), consistent with what we see in Table 5, above, resulting from teachers stretching out a weekend with a Monday or Friday absence in order to have a longer block of leisure.

[19] Attrition estimates predicting attrition from the school or from the state are reported in **Appendix B.**

$$\log\left(\frac{p_{jkdt}(y)}{1-p_{jkdt}(y)}\right) = \gamma_y + \gamma_1 T_{jt} + \gamma_2 S_{kt} + \gamma_3 SCREEN_j + \gamma_t + \varepsilon^{\gamma}_{ijst}, \quad (4)$$

The intercept in each level of this model, γ_y, is an indicator for the gap between t and $tprior$, as is present in the other models. The other control variables in Equation 4 are the same as the control variables in Equation 3, with the addition to the vector of teacher characteristics (T_{jt}) a series of indicators for whether the teacher holds an endorsement in a particular subject. The coefficient of interest in each level of Equation 4, γ_3, represents the expected change in the log odds of leaving the district correlated with a one-standard-deviation change in the teacher's screening score, all else being equal. Standard errors are clustered at the teacher level.

4.2 Correction for Sample Selection

The models described above estimate the extent to which the rubrics used as part of the hiring process predict the outcomes of interest. However, the findings from these models should not necessarily be interpreted as being *causal*. While we do observe all applicants to SPS, we only observe teacher outcomes for those teachers who are hired in a public school in Washington. This raises the concern that our findings could suffer from selection bias. For instance, teachers who are hired despite scoring poorly on the screening instruments may have been hired because they excel in areas not captured by the screening instruments. As Rockoff et al. (2011) observe, should these variables not captured by the instrument be positively correlated with, for instance, teacher effectiveness, the relationship between the instruments and effectiveness would be biased downward.

We note that there are several types of selection occurring in the case of teacher applicants to SPS. Applicants are selected by Spokane on the basis of their screening scores and other observable characteristics. Those who do not end up employed in Spokane may end up employed in public schools elsewhere in Washington or employed in another occupation in Washington, or they may leave the state altogether. The main model results include all Washington teachers who applied to Spokane; if no

teachers left the state or the occupation, the main model would not face selection bias. In this section, we model not the selection by the statewide group, observed in the main model, but the more specific selection by Spokane. Our approach to selection is intended to address the question of whether the main model results are likely to be biased by the 68.2% of applicants whom we do not observe teaching in a certificated position in Washington public schools during our sample window. If we find no bias, then this offers support for the main model findings.

Since the data set we use includes the entire applicant pool, we attempt to address the potential for selection bias by using variables that predict that an applicant will be hired to a job but do not otherwise predict teacher outcomes. Variables that satisfy these criteria are difficult to come by (e.g., Bound, Jaeger, & Baker, 1995; Staiger & Stock, 1997). We generate two variables designed to predict whether an applicant is likely to be hired but are otherwise uncorrelated with that applicant's quality as a teacher. [20]

1. A variable indicating whether an applicant was given a 60-point screening score because of the fact that he or she received a favorable tally error on the 21-point screening instrument.
2. A measure of the amount of competition faced by an applicant: the average 21-point screening scores of the other applicants for the job.

Tally errors arise from the incorrect hand-marking of 21-point screenings, which occurs in 18.8% of job applications and lead to the applicants' erroneously receiving 60-point screenings in 3.64% of job applications.[21] These errors may result from mistakes made in the addition of the sub-components, which account for about 38% of errors; forgetting to multiply the Recommendations criterion by 1.5,

[20] We also estimate the model, adding a measure of competition and errors in the totaling of the 60-point scores. However, errors in 60-point scores are not statistically significant in the first stage. In addition, this approach requires the first stage to be limited to those who received 60-point scores, which excludes an important part of the selection process. When this approach is used, second-stage results are similar.

[21] As we discussed above, applicants with missing 21-point components cannot be classified as having an incorrectly calculated score.

which appears to account for about 7% of errors; or remembering to multiply the Recommendations criterion but doing so incorrectly, which appears to account for 56% of errors.[22]

These rater errors can directly affect the decision on whether the applicant proceeds to the next stage of screening. In theory these rater errors should not be related to the quality of the teacher. The variable takes the form of an indicator for applicants whose calculated score is below the principal-requested cutoff for being advanced to the next stage but whose rater-added score is above the cutoff. In the case of jobs for which the cutoff score is not available (200 jobs), we count the number of applicants, N, who advance to the 60-point screening stage. The variable is, then, an indicator that the applicant's calculated 21-point score is not ranked among the top N calculated scores but whose rater-added score *is* among the top N rater-added scores.

Higher average rating scores among the competition should make it more difficult to land the job, and our assumption is that this should not predict a teacher's later performance. Justifying the omission of this variable from the model of teacher effectiveness does rely on the possibly troublesome assumption that a rater's scoring of an applicant is unaffected by the quality of the competition. To address this possibility, we perform tests of whether teacher performance residuals are related to the excluded variables.

We estimate the following teacher effectiveness model, using the instruments described above:

$$\hat{\alpha}_j|Hired = \alpha_0' + \alpha_1' SCREEN_{j(tprior)} + \varepsilon_{jst}^{\alpha'} \quad (5)$$

$$Hired^* = \alpha_{H0} + \alpha_{H1} Z_t + \alpha_{H2} SCREEN_{j(tprior)} + \varepsilon_t^H \quad (6)$$

$$Hired = I(Hired^* \geq 0) \quad (7)$$

[22] We cannot perfectly identify which type of error occurs in each case. We record the error as being the result of forgetting to multiply by 1.5 if the calculated score is greater than the rater score by .5 of the Letters of Recommendation score. We record as addition errors cases where the difference between the two scores is a whole number. Errors in multiplication are identified by differences between the two scores that are not a whole number.

Where *Hired** is the propensity for a particular applicant to be hired into Spokane on the basis of his or her 21-point screening score and the above-defined excluded variables Z_t. Equations 5 through 7 are estimated as a Heckman selection model, using a two-step method (Heckman, 1979; Maddala, 1983). We similarly estimate Heckman selection models for our other outcomes, absences, and attrition, combining Equations 3 and 4, in turn, with Equations 6 and 7. Given the assumptions about the exclusion restrictions, which are necessary to avoid collinearity between the selection correction term and the included variables, this generates OLS coefficients for Equation 4, which are corrected for selection into the sample.

Equation 6 (the first stage) is estimated at the job application level, which generates an estimated likelihood of being hired and thus being present in the sample, and shows the strength of the excluded predictors for leading to inclusion in the sample. In the second stage, the inverse Mills ratio, calculated using the linear prediction from the first stage, is included as a regressor in Equation 5,[23] which corrects for sample selection bias. Since the first stage predicts the probability of being hired by Spokane, the second-stage models include only applicants who are hired by the district during the observed sample window.

The first and second stages occur at different levels of observation: In the first stage, each observation is a single job application, but in the second stage, each observation is a teacher's value added, absences, or attrition in a particular year.[24] This problem is similar to that faced by Winters, Dixon, & Greene (2012), who point out that these issues make the estimation of standard errors on marginal effects in the second stage problematic. Our unadjusted standard errors are too small. We do not correct these standard errors, and thus our analysis is biased in favor of finding a statistically significant difference between corrected and uncorrected estimates. We argue that this is a reasonable

[23] As is standard in the use of the Heckman selection correction, the first stage is estimated using a probit model.
[24] Standard errors in the second stage are clustered at the teacher or teacher-year level, depending on outcome, but still do not take into account the difference in observation level between the first stage and second stage, as the first stage is at the application level and many teachers apply multiple times.

approach as long as we find no significant differences between corrected and uncorrected estimates, since the selection models are used in this paper largely as a robustness test to see if our primary results are biased by sample selection.

5. Predictive Validity of Screening Instruments

Below, we present the estimated coefficients from the primary models describing the relationship between the hiring rubrics used by SPS and: (a) teacher effectiveness (Section 5.1), (b) teacher absences (Section 5.2), and (c) teacher attrition (Section 5.3). In Section 5.4, we report findings from models that correct for sample selection.

The results below are estimated on the broadest sample available for each outcome. This includes teachers hired by Spokane during the hiring window detailed in our data, as well as teachers who go through Spokane's hiring process but end up employed elsewhere. While Chow tests suggest that it is appropriate to estimate the models separately for those employed in Spokane and those employed elsewhere, the coefficient estimates of interest (those on $SCREEN_{j(tprior)}$) are qualitatively similar for most outcomes when the sample is restricted to applicants employed by Spokane (these results are reported in **Appendix Table B3**). Results are stronger outside of Spokane for math achievement; this may be attributable to small sample sizes or heterogeneity in the sensitivity of the screening instrument along the performance distribution, a point we discuss further in Section 5.1. We also estimate the models separately, by school level (elementary, middle, and high), and find the 60-point score to predict more strongly for math and attrition in middle school. (These results are reported in **Appendix Table B4.)**[25]

[25] The student achievement models are not estimated at the high school level because, under Washington's testing regime, there are very few teachers for whom it is possible to calculate student achievement using a prior year's test score that is well-aligned with the outcome test score.

Prior to discussing the findings for the hiring rubrics, it is worth noting several unreported regression coefficients that are generally consistent with existing empirical literature. In the case of the first-step student achievement model (from Equation 1), for instance, we find that students eligible for free or reduced-price lunch score about 0.07 to 0.08 standard deviations lower than those who are not eligible.

We are also interested in the relationship between other teacher characteristics and teacher effectiveness; so we estimate alternative one-step models, similar to Equation 1 but in which teacher fixed effects are replaced by screening score controls and teacher characteristics. Experience has been shown to be an important predictor of achievement; thus, we estimate all the specifications in the table, with years of experience included in the model. Students assigned to first-year teachers relative to those assigned to second-year teachers score about 0.03 to 0.06 standard deviations lower on the state assessment, a finding similar to estimates from the literature (Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005; Clotfelter, Ladd, & Vigdor, 2006; Boyd, Lankford, Loeb, & Wyckoff, 2010; Goldhaber & Hansen, 2013). In our other outcomes, teachers in their first or second year are predicted to be absent about 1 day less often than teachers with 3 to 5 years of experience and almost 3 fewer days than teachers with 5 to 10 years of experience.

In addition, it seems logical that familiarity with the district will affect an applicant's degree of success in a job. This could be an argument for hiring internal transfers or applicants who did their student teaching in Spokane (and it is evident from **Table 5** that there is a strong preference for hiring applicants who did their student teaching in the district). We test this hypothesis by estimating one-step specifications that include an indicator for prior student teaching in Spokane and prior employment in the district in a certificated position. Neither coefficient is statistically significant.

5.1 Applicant Information and Student Achievement

The results of the predicted relationship between the 21- and 60-point rubrics and teacher effectiveness are presented in **Table 6**, both with and without school fixed effects. The rubric scores

have been normalized so that the coefficients should be interpreted as the effect of a one-standard-deviation change in an applicant's score (or, for Specifications 2 and 4, on one of the subcomponents that determine the rubric ratings) on the teacher fixed effect predicting student achievement in math (columns 1 and 2) and reading (columns 3 and 4).[26] With the exception of Specifications 5 and 6, which include either both screening scores (Specification 5) or both factors (Specification 6), respectively, each rubric rating or subcomponent coefficient is based on a model that includes only grade, year, and gap indicators, with one summative rating or one subcomponent. Thus, the coefficient on the rating is the effect of a standard deviation increase in that rating, holding constant other controls but not other screening score measures.[27]

Applicant scores on the 21-point rubric have a positive but insignificant relationship with teacher effectiveness in both math and reading (Specification 1, the top row in **Table 6**). Regarding the subcomponents of the 21-point rubric (Specification 2), it appears that the positive association between the 21-point rubric and teacher effectiveness is primarily related to the Recommendations component of the rubric, which has the largest coefficient for both subjects and is the component that is most heavily weighted in the construction of the 21-point rating (this component gets a 1.5 weight, while the Experience and Depth of Skills components get 1.0 weights).

The relationship between the 60-point rubric and teacher effectiveness (Specification 3) is greater than that of the 21-point score for both subjects and is statistically significant for math. This specification assumes a linear relationship between rubric scores and teacher effectiveness, but we see some evidence that the sensitivity of the rubric varies along the applicant performance distribution: In results available from the authors, we allow the effect of the screening scores to vary by the quartile of

[26] For reference, a standard deviation for the 21-point and 60-point rubric scores is approximately 2.4 and 7.3, respectively, and each subcomponent has a standard deviation of about 1.

[27] Thus, for instance, in the case of the models with the subcomponents, one *should not* consider the estimated coefficient on the reported subcomponent as the effect of a standard deviation change on that subcomponent holding constant other subcomponents, since the various subcomponents are positively correlated with one another (see **Table 4**).

the scores and find that the screening score is most informative in the top quartile for math but the second-to-bottom quartile for reading.

We suggest that both the 21- and 60-point results, including the nonsignificant results, predict improvements that are educationally meaningful.[28] Students assigned to teachers who score one standard deviation higher on the 60-point rubric are predicted to have student achievement that is 0.074 standard deviations higher in math and 0.033 standard deviations higher in reading. These effects are similar, for instance, to the estimated difference in achievement associated with being assigned to a novice teacher versus a second- or third-year teacher (0.03–0.06 standard deviations).[29]

Each of the subcomponents on the 60-point rubric receives an equal weight in the construction of the overall rating; yet there are substantial differences in their respective coefficient estimates (Specification 4). For instance, the coefficient on Classroom Management is relatively large for both math and reading; for reading, it is the only statistically significant coefficient.[30] Training, Flexibility, and Instructional Skills are also significant and large for math. The lack of significance for Certificate and Education is noteworthy, given the fact that certification is a measure to which many school systems give primacy (U.S. Department of Education, 1997). More generally, the fact that the estimated coefficients of the subcomponents of the 60-point rubric are substantially different from one another suggests that a reweighting of these subcomponents could increase the ability of the 60-point rubric rating to predict teacher effectiveness. We explore this issue more extensively in Section 6.

In Specification 6, we use standardized factors derived using factor analysis from all 13 screening components (3 from the 21-point score and 10 from the 60-point score) as predictors of student

[28] This specification assumes a linear effect of the screening scores. However, the screening scores may be more or less informative at different points across the rating performance distribution.

[29] Estimates for 60-point scores shrink when both 21- and 60-point scores are included simultaneously, which isn't surprising, given the positive correlation of the two rubrics.

[30] The Classroom Management finding is particularly interesting because inferences about an applicant's classroom management ability, like the Recommendations component of the 21-point rubric, are based on raters' assessments of the Letters of Recommendation. Moreover, classroom management has also been found to be an important predictor of student achievement in the context of classroom observation ratings of teachers (Tyler, Taylor, Kane, & Wooten, 2014).

outcomes. In keeping with the rest of the table, the 21-point factor is listed first, even though it is the second factor by the size of the eigenvalue. Coefficients on screening scores are similar to those in Specification 5.

Table 6 also reports estimates from models that include school fixed effects in the second step of the estimation. We estimate this specification, since a potential problem with the models that compare teachers across schools is their implicit presumption of equal rating standards across schools, a condition that may not hold.[31] However, since there may be some sorting of teacher quality across schools, models that include school fixed effects might mask some of the predictive power of the hiring rubrics.[32] It turns out, however, that the coefficients from the school fixed effects specifications are only slightly smaller than those estimated without school fixed effects (and not surprisingly, they are less precisely estimated).

5.2 Applicant Information and Teacher Absences

The results of the predicted relationship between the 21- and 60-point rubrics and teacher absences are presented in **Table 7**. Absences are measured in days so the coefficients show the estimated effect of a one-standard-deviation change in an applicant's screening score on the number of annual days a teacher is absent (columns 1 and 2), or the number of days that a teacher is absent on a Monday or Friday during a school year (columns 3 and 4). As in the previous section, each coefficient is from a separate regression, with the exception of Specifications 5 and 6.

The 21-point score is a positive predictor of teacher absences, whether it is entered into the model separately (Specification 1) or in tandem with the 60-point screening score (Specification 5), although

[31] Rating standards do appear to differ by school. We regress standardized 60-point scores on standardized 21-point scores and school fixed effects. An *F*-test on the school fixed effects is significant at the 99% level, and the fixed effects have a standard deviation of 0.448.

[32] District fixed effects are not included in the main model, since there are not enough teachers in most non-Spokane districts to identify these effects. Instead, an indicator for "in Spokane" is included. The school fixed effects model groups those who do not work in Spokane into a single "school."

the 21-point score is only significant when predicting Monday and Friday absences without school fixed effects—and even then, only at the 10% level. The point estimate suggests that a one-standard-deviation increase in screening score is predicted to *increase* teacher absences by about half of a day. The total 60-point screening score is insignificant in each specification, and few of the rubric subcomponents are significant, although Experience shows up as significant and large for yearly absences. The lack of a consistently significant relationship between screening scores and teacher absences is consistent with the null relationship between noncognitive skill and absences found in Rockoff et al. (2011).

Previous research has found a strong positive relationship between experience and teacher absences (Clotfelter, Ladd, & Vigdor, 2011; Herrmann & Rockoff, 2012), plausibly because teachers with experience are more likely to be tenured (Miller, Murnane, & Willet, 2008). We test whether the magnitude of the relationship between the rubric scores and teacher absences are related to the fact that the rubrics value experience by including it in the model. When we do this, the magnitudes of both summative ratings and specific components decrease substantially, with the 21-point score becoming insignificant and the 60-point score becoming slightly more negative, suggesting that the relationship between the rubric scores and teacher absences is indeed related to the fact that the rubrics reward experience.

5.3 Applicant Information and Teacher Attrition

Table 8 presents the relationship between standardized screening scores and the log odds of the probability that teachers leave their district, relative to staying in it. A positive coefficient indicates that a positive change in a particular variable increases the likelihood of attrition and a negative coefficient indicates that a positive change in that variable decreases the likelihood of attrition. In **Appendix Table B5**, we report findings on school and state attrition, which are very similar to findings reported on district attrition.

Scores on both the 21- and 60-point rubrics are predictive of decreased likelihood of district attrition. While the 21-point score is not statistically significant when school fixed effects are included, the magnitude of the coefficient changes very little. Applicants scoring higher on each of the two rubrics are less likely to leave their districts. These results are driven by a few of the subcomponents. On the 21-point rubric, it is the Depth of Skills component that is most strongly predictive of attrition. The 60-point components, as in mathematics achievement, have nonequal coefficients, and predictive power is centralized in a subset of the components: Experience, Classroom Management, Flexibility, Instructional Skills, Interpersonal Skills, and Preferred Qualifications all significantly predict less of each type of attrition. Effect sizes are in the range that a one-standard-deviation increase in the 60-point screening score is associated with about a 2.5 percentage point decrease in the probability of attrition. Given the fact that about 20% of hired teachers attrit after the first year (see **Table 5**), this is not a small change. **Figure 2**, which shows the estimated average cumulative attrition for teachers with rubric scores placing them in the bottom and top quartiles of the selection rubric rating distribution, provides a more concrete picture of whether the differences in performance are predictive of meaningful differences in teacher attrition. The differential in predicted district attrition for teachers with different scores on the 21-point rating ranges from 10 percentage points after 1 year to about 20 percentage points after 4 years.

5.4 Accounting for Sample Selection

The above findings suggest that the screening instruments are predictive of key teacher outcomes, but there is cause to be concerned that these findings could be biased by sample selection. Specifically, we only observe a narrow portion of the distribution of 21-point ratings—and presumably, of 60-point ratings (again, only those who perform well on the 21-point rating get a 60-point rating). Theory suggests this should bias results downward (Maddala, 1983; Rockoff et al., 2011), since those hired with low scores are likely to have impressive unobservable characteristics.

We account for the potential of sample selection by estimating Equations 6 and 7 with the second-stage models for each of the outcomes of interest. For these selection models to yield unbiased coefficients, the instruments we use to predict employment in Spokane— the rater errors on the 21-point score and the average 21-point screening scores of other applicants for the same job—must be reasonably correlated with the endogenous employment outcomes but must not otherwise belong in the models of teacher performance. We present the first-stage probit estimates of the selection model in **Table 9**. The level of observation here is at the job *application* level, and so there are many more observations here than in other analyses.

The excluded variables perform as expected in the first-stage model. An applicant is more likely to be hired if the rater's tally of his or her 21-point screening score is erroneously high, putting him or her above the cutoff for advancing to the next stage of the hiring process where he or she is given a 60-point score (column 1). An applicant who faces a higher level of competition for a particular job, as measured by the average 21-point screening score relative to other applicants, is significantly less likely to be hired (column 2).[33] The *F*-test on the excluded instruments is highly significant, suggesting sufficient exogenous variation in hiring to control for selection into the sample.

The third column of the table presents the results of a placebo test. Here we estimate whether an applicant's performance on Spokane's hiring rubrics, along with the instruments, predicts the likelihood of his or her being hired in another district in Washington when not hired in Spokane. While better performance on the 21-point rating instrument increases the likelihood that applicants will be hired by other districts—indicating that the factors that Spokane considers for hiring are consistent to some degree with those other districts are using to make hiring decisions—we do not find a significant relationship between the excluded variables and the likelihood of being employed in another district.

[33] Coefficient sizes are such that erroneously receiving a 60-point score increases the probability of being hired by 1.4 percentage points, and a 1-point increase in competition (on a 21-point scale) decreases the probability of being hired by 1.2 percentage points.

This lends support to our interpretation that the excluded variables directly affect the chances of being hired in Spokane rather than representing unobserved characteristics that might be correlated with both the probability of employment and our outcomes.

In order to be consistent with the selection process modeled in **Table 9**, only those who were hired for the job they applied for are included in the second stage. This is as opposed to the main model, which includes all teachers with observable outcomes, or the Spokane-only sample used in **Appendix Table B3,** which also includes teachers who applied for a transfer from one Spokane job to another and did not receive it. This means that the estimates in the second stage are made using a very restricted range of screening scores. The minimum 60-point screening score overall is 10, but the minimum among those who are hired is 31. The average of the 60-point screening scores among those who were hired is 6 points higher than the average among those who were not hired; on the 21-point screening, the average scores of those hired are 1.5 points higher than the scores of those not hired. Minimum licensure test scores (WEST-B) in reading and math scores are about two thirds of a standard deviation higher among applicants who were hired than among those who were not.

The selection correction offers an unbiased estimate of the effect of the screening score among those hired by Spokane.[34] However, the primary use of the screening score is to differentiate between teachers who are considered qualified to work in Spokane and those who are not. The effectiveness of the screening score may differ across its range, in which case the sample selection–corrected estimate is not the preferred estimate. We present selection-corrected results as a test for selection bias; if no selection bias is found, we suggest that the main model results are preferable.

We present second-stage estimates of the relationship between screening scores and outcomes in **Table 10**. Estimates are presented first without selection correction and then with the correction. The coefficients for student achievement are very similar, regardless of whether we correct for selection.

[34] By analogy, see Blundell and Costa Dias (2000), who note that, in the case of heterogeneous treatment effects, the Heckman model identifies the average treatment *on the treated*, and not the average treatment effect.

Attrition coefficients are somewhat larger in scale than main model coefficients, keeping in mind that these are linear coefficients, as compared with logit coefficients in the main model.[35] In each case, there is no statistically significant difference between selection-corrected and uncorrected estimates, and no meaningful difference except for absences. There are only small differences between corrected and uncorrected estimates, which suggests that the bias introduced by the process of sample selection does not greatly affect estimates in the main model.[36,37] The relationship between the subcomponents of the hiring rubrics and the three teacher outcomes is not reported, but we also find no statistically significant change in the estimates of that relationship when we correct for sample selection.

For each model, we test the plausibility of the exclusion restrictions by estimating a pseudo-Sargan test for overidentification. Residuals from the second stage are regressed on the set of excluded variables, and the *F*-test of all excluded variables is taken as a rough test on the exogeneity of the excluded variables. With the exception of reading achievement, we cannot reject the hypothesis that the residuals are unrelated to the excluded variables except through selection into the sample, buttressing the findings from the placebo test.

[35] The use of a linear probability model in the second stage is not technically the right specification, as this specification violates the Heckman model assumption of normally distributed errors. However, we believe it is appropriate for the purpose of tracking the change in the coefficient when the Mills ratio is added to the equation.

[36] As mentioned in Section 4.2, standard errors in the selection-corrected models are too low, and thus our results are biased in favor of finding statistically significant differences between selection-corrected and selection-uncorrected estimates. We find no differences, and these results represent an upper bound on the number of differences.

[37] We examine the difference in coefficient sizes rather than the statistical significance of the coefficient on the Mills ratio, since tests of the significance of the Mills ratio are underpowered because of collinearity. However, the finding that differences are "small" is not well defined. We perform a power analysis to determine the probability that the model finds a difference of a given size. We simulate all data in the model, matching the means and standard deviations of the screening scores, excluded variables, and student achievement. We estimate the model, using 1,000 sets of simulated data with and without the Heckman correction, and record the size of the difference in coefficients on screening scores. The model has 80% power to detect a difference of 0.064 or larger in the student achievement model. As long as our presentation of a "meaningfully large difference" is smaller than 0.064, we have reasonable power to find meaningfully large differences, which we do not.

6. Policy Implications and Conclusions

Our findings show that the two screening rubrics used by Spokane Public Schools predict teacher effectiveness and teacher attrition—but not teacher absences—in expected ways. For some perspective on this, a one-standard-deviation increase in the 60-point screening score is associated with about a 0.07 standard deviation increase in math achievement, a marginally significant (0.03 to 0.05) increase in reading achievement, and a decrease in attrition by about 2.5 percentage points.[38] Since the turnover of a single teacher can cost a district in the region of $10,000 (Barnes, Crowe, & Schaefer, 2007), improved hiring practices have the ability to both save money and improve effectiveness.[39]

The screening scores used by SPS represent the value of guided human interpretation of somewhat subjective information, such as that contained in letters of recommendation, and our findings validate the notion that this type of guidance on the way to interpret applications is an improvement on the ad hoc hiring processes typically seen in public schools (e.g., Ebmeier & Ng, 2006; Oyer & Schaefer, 2011).

To get a clearer sense of the value of more nuanced assessments, we use a factor analysis of objective criteria to illustrate that the screening scores provide predictive power above and beyond what could be achieved by making hiring decisions on the basis of objective, observable factors alone. In **Appendix Table B6,** we show the predictive validity of factors derived from objective criteria, with and without screening score controls. In the case of value added, for instance, the student-level standard deviation in teacher quality predicted by the purely objective factors is 0.085 for math and 0.063 for reading, as compared with state-wide standard deviations in teacher effects of 0.19 for math and 0.18

[38] The result that the screening scores are more significantly related to mathematics achievement than to reading achievement is consistent with prior findings (Dobbie, 2011; Rockoff et al., 2011).

[39] Attrition results are robust over different approaches to estimation. Teacher effectiveness results are somewhat sensitive to the inclusion of school fixed effects, but even with the inclusion of fixed effects, the magnitudes of coefficients on the rubric scores remain comparable (though not statistically significant) to the differential between novice and second-year teachers.

for reading (Goldhaber & Theobald, 2013).[40] This implies that the objective information obtained through the application process explains about 20% of the variance in teacher effectiveness in math and 12% in reading. With the addition of screening scores, the estimated standard deviation in the measured teacher effect increases to 0.100 for math and 0.067 for reading, so that the explained portion of teacher effectiveness rises to 28% in math (a 40% increase) and 14% in reading (a 17% increase).[41]

While the hiring rubrics appear to be effective, analysis of the subcomponents suggests that Spokane could increase the predictive validity of the summative rubric ratings by reweighting the components. To illustrate this, we estimate the canonical correlation between each of the various subcomponents of the 60-point rubric and each of the teacher outcomes in order to derive the weights that optimize the correlation between the rubric rating and the outcomes.[42,43]

Table 11 presents the optimal weights (recall that each component has an equal weight of 0.1 under the existing 60-point rubric). There is a great deal of heterogeneity in the optimal weights across the different outcomes. This may suggest that the rubric is not identifying underlying traits that are equally predictive of different types of achievement.

The weights that maximize the outcomes are different for each outcome. The only component that is given more than a 0.1 weight in every outcome is Certificate and Education; while, by itself, the Certificate and Education subcomponent is not a strong predictor of teacher effectiveness, this subcomponent is not strongly correlated with the others and so enters positively here because it provides some measure of information beyond the other factors. There are, however, several subcomponents that could be significantly and consistently down-weighted across all different

[40] These are from models that do not include school fixed effects.
[41] These results can be compared to those of Rockoff et al. (2011), who find that the introduction of nontraditional information collected about applicants increases the standard deviation in the measure of teacher effectiveness by 50%, so that in total about 10% of the variance in teacher effectiveness is explained.
[42] We do not include the 21-point rubric in this weighting, since the district has since abandoned the 21-point rubric for a 42-point rubric; thus, any weighting of the components in the 21-point rubric would not be useful.
[43] The weights were all constrained to be non-negative and to sum to 1.

outcomes; this suggests that some aspects of the hiring rubric, or the training in how to implement it, could be improved. These results also suggest the capacity to improve teacher training. If teacher training is able to focus on improving the skills found here to relate strongly to teacher effectiveness, then incoming teachers will be more capable of improving student learning.

To provide a sense of the potential gains from reweighting the 60-point rubric, we estimate the predicted coefficient on a reweighted 60-point rubric for each of the outcomes and, by way of comparison, also show the coefficient estimated with equal component weighting. The results of this exercise are at the bottom of in **Table 11;** not surprisingly, there is a large increase in the magnitude of the reweighted 60-point rubric, especially in the math model (column 1) and for attrition (column 4).[44] There are, however, at least two cautions that should be applied to interpreting the above findings from **Table 11.** First, we do not know whether some of the components that do a poor job of predicting the outcomes studied here might predict other teacher behaviors or student outcomes that are valued by school districts but not measured.

Second, and more generally, our findings for Spokane may not generalize across all school districts. Spokane is seen as a desirable place to work in eastern Washington and may not face the same hiring problems as other districts, which may face more competition for teacher labor. Spokane also hires a high percentage of its workforce from among people who already have experience there. Slightly more than 70% of jobs were filled with applicants who either already worked in Spokane or had been student teachers there. It is possible that the predictive abilities of the screening rubrics are aided by the fact that the screeners may be familiar with those who are writing the letters of recommendation. Nevertheless, these results are consistent with the wider literature on screening at the hiring stage in other industries, as well as in teaching.

[44] Even with the components optimally weighted, while the coefficient on the 60-point screening score increases, it still does not significantly predict teacher absences or reading achievement.

The idea of improving the quality of the teacher workforce through more effective hiring is appealing, given the high-dollar and political costs of dismissing ineffective teachers who are in service (Treu, 2014), and empirical evidence that finds that other teacher performance interventions, such as professional development or performance incentives, tend to have marginal impact on productivity. The evidence we present here shows a strong relationship between the performance on selection instruments and some measures of in-service teacher quality. This relationship likely overstates what is possible in terms of improving the teacher workforce as a whole, since school systems compete with one another in the market for teacher labor. Nevertheless, since many school districts rely on far more informal processes for selecting teacher, and likely lose some potentially talented teachers to other occupations at the hiring stage, there appears to be substantial room for improving the quality of the teacher workforce through greater use and refinement of teacher selection instruments.

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.

Ballou, D. (1996). Do public schools hire the best applicants? *The Quarterly Journal of Economics*, *111*(1), 97–133.

Barnes, G., Crowe, E., & Schaefer, B. (2007). *The cost of teacher turnover in five school districts: A pilot study.* Washington, DC: National Commission on Teaching and America's Future.

Bliesener, T. (1996). Methodological moderators in validating biographical data in personnel selection. *Journal of Occupational and Organizational Psychology, 69*(1), 107–120.

Blundell, R., & Costa Dias, M. (2000, December). Evaluation methods for non-experimental data. *Fiscal Studies, 21*(4), 427–468.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995, January). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

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.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement. *Education Finance and Policy, 1*(2), 176–216.

Boyd, D. J., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008). *Measuring effect sizes: The effect of measurement error*. Madison, WI: National Conference on Value-Added Modeling. University of Wisconsin–Madison.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440.

Boyd, D. J., Lankford, H., Loeb, S., & Wyckoff, J. H. (2010). *Teacher Layoffs: An empirical illustration of seniority vs. measures of effectiveness* (Brief 12). Washingon, DC: National Center for Analysis of Longitudinal Data in Education Research.

Boyd, D. J., Lankford, H., Loeb, S., & Wyckoff, J. (2013). Analyzing the determinants of the matching of public school teachers to jobs: Disentangling the preferences of teachers and employers. *Journal of Labor Economics*, *31*(1), 83–117.Cannata, M., Rubin, M., Goldring,

E., Grissom, J. A., Neumerski, C., Drake, T., & Schuermann, P. (2014, March). *Using teacher effectiveness data for information rich hiring* (p. 47). Paper presented at the annual meeting of the Association for Education Finance and Policy, San Antonio, TX .

Chamberlain, G. E. (2013). Predictive effects of teachers and schools on test scores, college attendance, and earnings. *Proceedings of the National Academy of Sciences of the United States of America, 110*(43), 17176–17182.

Chetty, R., Friedman, J., & Hilger, N. (2010). *How does your kindergarten classroom affect your earnings? Evidence from Project STAR* (NBER Working Paper No. 16381). Cambridge, MA: National Bureau of Economic Research.

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, 24*(9), 2593–2632.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778–820.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2009). Are teacher absences worth worrying about in the United States? *Education Finance and Policy, 4*(2), 115–149.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2011). Teacher mobility, school segregation, and pay-based policies to level the playing field. *Education Finance and Policy, 6*(3), 399–438.

Dee, T., & Wyckoff, J. (2013). *Incentives, selection, and teacher performance: Evidence from IMPACT* (NBER Working Paper No. 19529). Cambridge, MA: National Bureau of Economic Research.

Dobbie, W. (2011). *Teacher characteristics and student achievement: Evidence from Teach for America*. Cambridge, MA: Harvard University.

Ebmeier, H., & Ng, J. (2006). Development and field test of an employment selection instrument for teachers in urban school districts. *Journal of Personnel Evaluation in Education, 18*(3), 201–218.

Figlio, D. N., & Kenny, L. W. (2007). Individual teacher incentives and student performance. *Journal of Public Economics, 91*(5–6), 901–914.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Education Research Journal, 38*(4), 915–945.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Walters, K., . . . Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation*. Washington, DC: Institute for Education Sciences.

Glazerman, S., Mayer, D. P., & Decker, P. T. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75–96.

Goldhaber, D. (2013). *What do value-added measures of teacher preparation programs tell us?* (Knowledge brief 12). Carnegie Knowledge Network.

Goldhaber, D., & Anthony, E. (2003). *Teacher quality and student achievement*. *Urban diversity series*. New York: ERIC Clearinghouse on Urban Education.

Goldhaber, D., & Brewer, D. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis*, *22*(2), 129–145.

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica*, *80*(319), 589–612.

Goldhaber, D., Krieg, J., Theobald, R., & Brown, N. (2014). The STEM and special education teacher pipelines: Why don't we see better alignment between supply and demand? (CEDR Working Paper 2014-3). Seattle, WA: University of Washington.

Goldhaber, D., & Theobald, R. (2013). Managing the teacher workforce in austere times: The determinants and implications of teacher layoffs. *Education Finance and Policy*, *8*(4), 494–527.

Goldhaber, D., & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp Teacher Pay Initiative. *Economics of Education Review*, *31*(6), 1067–1083.

Guthrie, J. W., & Rothstein, R. (1999). Enabling "adequacy" to achieve reality: Translating adequacy into state school finance distribution arrangements. In H. F. Ladd, R. Chalk, & J. S. Hansen (Eds.), *Equity and adequacy in education finance: Issues and perspectives* (pp. 202–259). Washington, DC: National Academy Press.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy*, *100*(1), 84–117.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality* (NBER Working Paper No. 11145). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267–271.

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.

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

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153–161.

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

Herrmann, M. A., & Rockoff, J. E. (2012). Worker absence and productivity: Evidence from teaching. *Journal of Labor Economics, 30*(4), 749–782.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*(4), 430–511.

Hill, H. C., & Grossman, P. (2013). Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371–385.

Hinrichs, P. (2013). *What kind of teachers are schools looking for? Evidence from a randomized field experiment.* Paper presented at the 38th Annual AEFP Conference, New Orleans, LA.

Ingersoll, R. M., & Perda, D. (2010). Is the supply of mathematics and science teachers sufficient? *American Educational Research Journal, 47*(3), 563–594.

Jackson, C. K. (2012). *Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina* (NBER Working Paper No. 18624). Cambridge, MA: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics, 64*(3), 560–572.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. New York: Cambridge University Press.

Mason, R. W., & Schroeder, M. P. (2010). Principal hiring practices: Toward a reduction of uncertainty. *The Clearing House, 83*(5), 186–193.

McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41*(2), 283–309.

Metzger, S. A., & Wu, M.-J. (2008). Commercial teacher selection instruments: The validity of selecting teachers through beliefs, attitudes, and values. *Review of Educational Research, 78*(4), 921–940.

Miller, R. T., Murnane, R. J., & Willet, J. B. (2008). Do teacher absences impact student achievement? Longitudinal evidence from one urban school district. *Educational Evaluation and Policy Analysis, 30*(2), 181–200.

National Council on Teacher Quality. (2014). *2013 state teacher policy yearbook: National summary*.

Neal, D. (2011). *The design of performance pay in education* (NBER Working Paper No. 16710). Cambridge, MA: National Bureau of Economic Research.

Oyer, P., & Schaefer, S. (2011). Personnel economics: Hiring and incentives. In D. Car & O. Ashenfelter (Eds.), *Handbook of labor economics* (Vol. 4., pp. 1769–1823). Elsevier.

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

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

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43–74.

Ronfeldt, M., Loeb, S., & Wyckoff, J. H. (2013). How teacher turnover harms student achievement. *American Educational Research Journal, 50*(1), 4–36.

Shaw, K., & Lazear, E. (2007). Personnel economics: The economist's view of human resources. *Journal of Economic Perspectives, 21*(4), 91–114.

Society for Industrial and Organizational Psychology. (2014). Employment Testing. Retrieved from http://www.siop.org/workplace/employment%20testing/employment_testing_toc.aspx

Springer, M. G., Ballou, D., Hamilton, L. S., Le, V.-N., Lockwood, J., McCaffrey, D. F., . . . Stecher, B. M. (2010). *Teacher pay for performance: Experimental evidence from the Project on Incentives in teaching*. Nashville, TN: RAND Corporation.

Staiger, D. O., & Rockoff, J. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives, 24*(3), 97–118.

Staiger, D. O., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586.

Strauss, R. P., Bowes, L. R., Marks, M. S., & Plesko, M. R. (2000). Improving teacher preparation and selection: Lessons from the Pennsylvania experience. *Economics of Education Review*, *19*(4), 387–415.

Treu, R. M. *Vergara vs. State of California Tentative Decision* (2014).

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2014). Using student performance data to identify effective classroom practices. *The American Economic Review, 100*(2), 256–260.

U.S. Department of Education. (1997). *Credentials and tests in teacher hiring: What do districts require?* Washington, DC: Author.

Winters, M. A., Dixon, B. L., & Greene, J. P. (2012). Observed characteristics and teacher quality: Impacts of sample selection on a value added model. *Economics of Education Review*, *31*(1), 19–32.

Xu, Z., Hannaway, J., & Taylor, C. (2011). Making a difference? The effects of Teach for America in high school. *Journal of Policy Analysis and Management, 30*(3), 447–469.

Tables and Figures

Table 1. Evaluation Criteria on 21-Point Prescreening Rubric

Criterion	Screener should look for...
Experience related to position	Years of experience, type of experience, type of school/district, gaps in teaching experience
Depth of Skills	Evidence of strong content knowledge, strong classroom management, differentiates instruction, engages parents/families, strong rapport with students and colleagues, commitment to the school as a community, socially just practices, experience with diverse student populations, makes learning relevant, engages students in active learning, elementary level currently seeking those with experience using Fosnot, Calkins, GLAD strategies, response to intervention
Quality of recommendations	All items noted in above categories, does writer recommend/strongly recommend, personal or professional recommendation, does the writer regularly evaluate teachers (preference of letter from principal, asst. principal, instructional coach, supt.)

Table 2. Evaluation Criteria on 60-Point Screening Rubric

Criterion	Screener should look for...
Certificate and Education	Note completion of course of study; certificate held (current or pending); education.
Training	Look for quality, depth, and level of candidate's additional training related to position.
Experience	Note degree to which experience supports the prediction of success, not just the number of years. A beginning candidate could be rated highly.
Classroom Management	Look for specific references to successful strategies. This may not mean quiet and orderly, but planned and directed. Effectively handles large/small or ethnically/sociologically diverse groups; develops routines and procedures to increase learning, establishes clear parameters, and responds appropriately.
Flexibility	Note multiple endorsements, activity, coaching interests, student, building or district, or community support. Willing to learn new concepts and procedures, successfully teaches a variety of assignments, effectively uses various teaching styles.
Instructional Skills	Look for specific references in support of skill in this area – plans, implements, evaluates, relates to students, creative, multiple approaches, monitors and adjusts, uses culturally responsive strategies appropriate to age, background and intended learning of students.
Interpersonal Skills	Develops and maintains effective working relationships with diverse staff, students, parents/guardians, and community.
Cultural Competency	Look for specific references to successful strategies for building and maintaining a relationship with each student and their family. This may not be explicitly mentioned, but the following strategies offer some evidence of cultural competency: specific instructional strategies providing each student access to a rigorous curriculum, inclusive/respectful language about students and families, a belief that all children can achieve at high levels, mention of conflict resolution/restorative practices, specific instructional strategies for integrating culturally responsive materials which are also rigorous, and appropriate statements about their work with diverse populations. Note relevant training, course work, authors/book titles listed.
Preferred Qualifications	Applicant has preferred qualifications as indicated in the job posting.
Letters of Recommendation	Look for current letters of recommendation from the most recent supervisor(s). Your score should reflect the quality and recentness of the recommendation, as well as the author of the letter.

Table 3. Applicant Screening Scores: Descriptive Statistics

		Unadjusted					Adjusted				
		Obs.	Mean	SD	Min.	Max.	Obs.	Mean	SD	Min.	Max.
21-Point Pre-Screening Rubric											
Rater Total Rating		3,944	16.1	(2.4)	4.5	21	3,944	16.1	(2.4)	4.5	21
Calculated Total Rating		2,614	15.9	(2.4)	3.5	21	2,672	16.0	(2.4)	3.5	21
21-Point Components	Experience	2,616	4.4	(0.8)	1	6	2,672	4.5	(0.8)	1	6
	Depth of Skills	2,616	4.7	(0.9)	1	6	2,672	4.8	(0.8)	1	6
	Recommendations	2,614	4.5	(0.9)	1	6	2,672	4.5	(0.9)	1	6
60-Point Screening Rubric											
Rater Total Rating		1,697	37.9	(7.6)	10	66	1,711	41.3	(7.3)	10	66
Calculated Total Rating		1,709	37.9	(7.5)	10	59	1,711	41.4	(7.3)	10	60
60-Point Components	Certificate and Education	1,673	5.1	(1.0)	0	6	1,711	5.0	(1.0)	0	6
	Training	1,704	3.9	(1.2)	0	6	1,711	3.9	(1.2)	0	6
	Experience	1,708	4.0	(1.1)	0	6	1,711	4.0	(1.1)	0	6
	Management	1,702	4.1	(1.0)	0	6	1,711	4.0	(1.1)	0	6
	Flexibility	1,705	4.2	(1.0)	0	6	1,711	4.2	(1.0)	0	6
	Instructional Skills	1,708	4.1	(1.0)	0	6	1,711	4.1	(1.0)	0	6
	Interpersonal Skills	1,705	4.4	(1.0)	0	6	1,711	4.4	(1.0)	0	6
	Cultural Competency	1,704	4.0	(1.0)	0	6	1,711	4.0	(1.0)	0	6
	Preferred Qualifications	1,472	3.9	(1.3)	0	6	1,711	3.6	(1.4)	0	6
	Letters of Rec.	717	4.1	(1.1)	0	6	1,711	4.1	(0.8)	0	6

Table 4. Pair-Wise Correlations of Applicant Screening Scores

		21-Point Pre-Screening Rubric				60-Point Screening Rubric										
		Total	Experience	Depth of Skills	Recommendations	Total	Certificate & Education	Training	Experience	Classroom Mgmt.	Flexibility	Instructional Skill	Interpersonal Skill	Cultural Competency	Preferred Quals.	Letters of Rec.
	21-Point Pre-Screening Rubric															
21-Point Components	Total Summative Rating	1.00														
	Experience	0.56	1.00													
	Depth of Skills	0.82	0.37	1.00												
	Recommendations	0.85	0.22	0.71	1.00											
	60-Point Screening Rubric															
60-Point Components	Total Summative Rating	0.17	0.13	0.17	0.10	1.00										
	Certificate & Edu.	0.03	0.03	0.01	0.02	0.38	1.00									
	Training	0.13	0.14	0.11	0.07	0.69	0.26	1.00								
	Experience	0.24	0.28	0.12	0.11	0.70	0.29	0.64	1.00							
	Classrm. Mgmt.	0.22	0.09	0.24	0.18	0.73	0.23	0.44	0.50	1.00						
	Flexibility	0.17	0.07	0.19	0.13	0.75	0.24	0.47	0.52	0.71	1.00					
	Instructional Skill	0.23	0.12	0.23	0.16	0.78	0.22	0.56	0.60	0.74	0.69	1.00				
	Interpersonal Skill	0.20	0.08	0.21	0.15	0.74	0.25	0.47	0.51	0.67	0.76	0.69	1.00			
	Cultural Comp.	0.12	0.07	0.11	0.10	0.65	0.16	0.48	0.46	0.51	0.55	0.53	0.56	1.00		
	Preferred Qual.	0.05	0.02	0.06	0.03	0.68	0.28	0.50	0.51	0.42	0.45	0.52	0.42	0.39	1.00	
	Letters of Rec.	0.20	0.06	0.22	0.19	0.73	0.19	0.42	0.51	0.69	0.63	0.67	0.63	0.54	0.48	1.00

Table 5. Outcome Variable Summary Statistics

	All	21-Pt Pre-Screening Summ. Rating	60-Pt Screening Summ. Rating	Interview	Hired/ Offered	Hired Elsewhere
Total Obs. (Teacher/Yr.)	4,217	3,944	1,709	1,238	538	498
Total Proportions	1.00	0.94	0.41	0.29	0.13	0.12
Applicant Information						
Certificated Employment Experience in Year Applied						
No Experience	0.83	0.84	0.68	0.63	0.49	0.53
	(0.38)	(0.36)	(0.47)	(0.48)	(0.50)	(0.50)
SPS District	0.11	0.09	0.22	0.28	0.43	0.03
	(0.31)	(0.28)	(0.42)	(0.45)	(0.49)	(0.17)
Other District	0.07	0.07	0.09	0.09	0.08	0.44
	(0.25)	(0.25)	(0.29)	(0.29)	(0.27)	(0.50)
Calculated Experience	3.18	3.23	3.87	3.73	3.24	4.43
	(4.66)	(4.64)	(5.02)	(4.74)	(4.23)	(5.30)
Student Teaching in SPS? (Y/N)	0.36	0.37	0.40	0.42	0.47	0.29
	(0.48)	(0.48)	(0.49)	(0.49)	(0.50)	(0.46)
21-Point Pre-Screening Rubric Summative Rating	NA	16.10	16.99	17.13	17.27	16.49
		(2.36)	(2.21)	(2.19)	(2.16)	(2.22)
60-Point Screening Rubric Summative Rating	NA	NA	41.34	43.62	45.66	40.19
			(7.32)	(6.19)	(5.74)	(7.06)
WEST-B Average	-0.07	-0.07	-0.03	-0.02	0.02	-0.04
(Standardized statewide)	(0.75)	(0.75)	(0.75)	(0.75)	(0.70)	(0.75)
(N = 1364 Teachers)						
Outcomes*						
Value-Added						
Math	-0.05	-0.06	-0.03	-0.03	-0.01	-0.08
(N=348 Teacher/Yr.)	(0.21)	(0.21)	(0.21)	(0.21)	(0.21)	(0.19)
Reading	-0.08	-0.09	-0.08	-0.07	-0.07	-0.09
(N=364 Teacher/Yr.)	(0.17)	(0.17)	(0.17)	(0.18)	(0.18)	(0.17)
Absences (N=1057 Teacher/Yr.)						
Total Annual Absences	6.92	6.62	7.38	7.51	7.27	5.28
	(5.35)	(5.09)	(5.24)	(5.32)	(5.33)	(5.10)
Total Monday/Friday Absences	3.12	2.98	3.33	3.37	3.29	2.44
	(2.50)	(2.42)	(2.51)	(2.50)	(2.48)	(2.47)
Attrit within 1 Year (N=1020 Teacher/Yr.)						
School	0.46	0.46	0.46	0.44	0.40	0.47
	(0.50)	(.50)	(0.50)	(0.50)	(0.49)	(0.50)
District	0.30	0.31	0.29	0.27	0.20	0.39
	(0.46)	(0.46)	(0.45)	(0.44)	(0.40)	(0.49)
K-12 WA Public Schools	0.22	0.23	0.22	0.21	0.17	0.17
	(0.41)	(0.42)	(0.41)	(0.41)	(0.37)	(0.38)
Attrit within 3 Years (N=780 Teacher/Yr.)						
School	0.46	0.46	0.45	0.43	0.40	0.51
	(0.50)	(0.50)	(0.50)	(0.50)	(0.49)	(0.50)
District	0.31	0.33	0.30	0.29	0.21	0.42
	(0.46)	(0.47)	(0.46)	(0.45)	(0.41)	(0.50)
K-12 WA Public Schools	0.22	0.23	0.22	0.22	0.17	0.17
	(0.42)	(0.42)	(0.41)	(0.41)	(0.38)	(0.38)

No experience, experience in SPS and experience in other districts determined by identifying applicants as being employed in a certificated teaching position. Value-added scores are estimated as a derivative of Equation 1 on page 12. WEST-B scores are centered at mean zero at the state level with standard deviations of approximately 0.20 and 0.16 for math and reading respectively (depending on year). *Observation numbers in the Outcomes panel represent the number of applications (at the teacher/year level) with associated outcome data. The numbers of observations of observed teacher/year outcome data are smaller, and are shown (conditional on having observed screening scores) as the number of clusters in each regression in **Tables 6-8**.

Table 6. Predictors of Teacher Effectiveness

	Math		Reading	
		School FE		School FE
(Spec. 1) 21-Point Score	N = 222 (185)[a]		N = 229 (189)	
21-Point Score	0.036	0.030	0.024	0.016
	(0.022)	(0.019)	(0.015)	(0.018)
(Spec. 2) 21-Point Components[b]	N = 151 (126)		N = 144 (123)	
Experience	0.033	0.043	0.023	0.024
	(0.028)	(0.029)	(0.019)	(0.023)
Depth of Skills	0.040	0.062**	0.004	-0.008
	(0.026)	(0.030)	(0.019)	(0.029)
Recommendations	0.059*	0.049	0.036	0.031
	(0.030)	(0.031)	(0.022)	(0.028)
(Spec. 3) 60-Point Score	N = 154 (128)		N = 151 (126)	
60-Point Score	0.074**	0.065*	0.033	0.048*
	(0.028)	(0.036)	(0.025)	(0.029)
(Spec. 4) 60-Point Components	N = 154 (128)		N = 151 (126)	
Certificate & Education	0.028	0.015	-0.000	0.023
	(0.040)	(0.053)	(0.029)	(0.040)
Training	0.069**	0.053	0.040	0.049
	(0.030)	(0.034)	(0.025)	(0.031)
Experience	0.045	0.003	0.010	0.004
	(0.034)	(0.039)	(0.027)	(0.031)
Classroom Management	0.133**	0.101**	0.043*	0.046
	(0.032)	(0.028)	(0.026)	(0.030)
Flexibility	0.091**	0.080**	0.032	0.052*
	(0.032)	(0.033)	(0.029)	(0.031)
Instructional Skills	0.063*	0.035	0.033	0.030
	(0.033)	(0.035)	(0.026)	(0.031)
Interpersonal Skills	0.043	0.059*	0.010	0.014
	(0.037)	(0.031)	(0.028)	(0.028)
Cultural Competency	0.018	0.028	-0.004	-0.006
	(0.026)	(0.031)	(0.023)	(0.027)
Preferred Qualifications	0.034	0.022	0.041	0.023
	(0.032)	(0.034)	(0.026)	(0.040)
Letters of Recommendation	-0.046	0.016	-0.070**	0.049
	(0.050)	(0.046)	(0.023)	(0.031)
(Spec. 5) 21-*and* 60-Point Scores	N = 132 (107)		N = 128 (104)	
21-Point Score	0.016	0.015	0.029	0.016
	(0.024)	(0.027)	(0.021)	(0.028)
60-Point Score	0.048	0.044	0.003	0.010
	(0.035)	(0.052)	(0.032)	(0.044)
(Spec. 6) Factor Analysis	N = 95 (77)		N = 90 (76)	
Factor 2 – 21-Point Score	0.013	0.035	0.010	-0.007
	(0.026)	(0.032)	(0.022)	(0.036)
Factor 1 – 60-Point Score	0.032	0.040	0.028	0.005
	(0.027)	(0.042)	(0.027)	(0.038)

Notes: Each of the first-step specifications includes controls for prior student test scores in math and reading, and a vector of student-level controls (gender, ethnicity, learning disability status, gifted program status, and free-or-reduced-lunch status). Each of the second-step specifications includes grade, year, and gap indicators. Standard errors are clustered at the teacher level. In each case the first-stage R^2 value is approximately 0.6, and the second-stage R^2 value without school fixed effects is approximately 0.1. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.

[b]With the exception of Specifications 5 and 6, each coefficient is estimated in a separate regression.

Table 7. Predictors of Teacher Absences (Sick Days Taken)

	Yearly Absences	School FE	Monday/Friday Absences	School FE
(Spec. 1) 21-Point Score	N = 453 (335) [a]		N = 453 (335)	
21-Point Score	0.439	0.222	0.266*	0.219
	(0.298)	(0.334)	(0.148)	(0.158)
(Spec. 2) 21-Point Components[b]	N = 304 (231)		N = 304 (231)	
Experience	0.356	0.245	0.107	0.080
	(0.338)	(0.375)	(0.166)	(0.185)
Depth of Skills	-0.292	-0.312	-0.100	-0.155
	(0.440)	(0.519)	(0.204)	(0.247)
Recommendations	-0.079	-0.232	-0.019	-0.058
	(0.362)	(0.426)	(0.168)	(0.191)
(Spec. 3) 60-Point Score	N = 287 (213)		N = 287 (213)	
60-Point Score	-0.054	-0.215	0.044	0.158
	(0.503)	(0.618)	(0.275)	(0.306)
(Spec. 4) 60-Point Components	N = 287 (213)		N = 287 (213)	
Certificate & Education	0.213	0.207	0.350	0.265
	(0.537)	(0.552)	(0.267)	(0.294)
Training	0.130	0.133	0.020	0.199
	(0.529)	(0.537)	(0.297)	(0.296)
Experience	1.121**	1.170**	0.355	0.559*
	(0.447)	(0.554)	(0.269)	(0.307)
Classroom Management	-0.267	-0.448	-0.044	-0.022
	(0.487)	(0.551)	(0.230)	(0.268)
Flexibility	-0.132	-0.326	-0.085	0.065
	(0.599)	(0.791)	(0.309)	(0.328)
Instructional Skills	-0.360	-0.644	-0.041	0.023
	(0.547)	(0.754)	(0.249)	(0.330)
Interpersonal Skills	-0.508	-0.502	-0.093	0.057
	(0.454)	(0.585)	(0.250)	(0.298)
Cultural Competency	0.027	-0.241	-0.242	-0.372
	(0.483)	(0.663)	(0.235)	(0.337)
Preferred Qualifications	0.342	0.387	0.307	0.605*
	(0.628)	(0.638)	(0.289)	(0.308)
Letters of Recommendation	-0.230	-0.103	-0.087	-0.139
	(0.416)	(0.544)	(0.260)	(0.319)
(Spec. 5) 21- and 60-Point Scores	N = 272 (205)		N = 272 (205)	
21-Point Score	0.486	-0.081	0.238	0.026
	(0.702)	(0.774)	(0.321)	(0.359)
60-Point Score	-0.091	-0.067	-0.010	0.167
	(0.523)	(0.653)	(0.288)	(0.339)
(Spec. 6) Factor Analysis	N = 198 (155)		N = 198 (155)	
Factor 2 – 21-Point Score	-0.581	-0.744	-0.288	-0.398
	(0.486)	(0.630)	(0.243)	(0.337)
Factor 1 – 60-Point Score	-0.348	0.144	-0.023	0.298
	(0.510)	(0.663)	(0.260)	(0.363)

Notes: Each specification controls for gender, ethnicity, school size, school percentages for students eligible for free/reduced lunch and for under-represented minorities, and indicators for school level, Title I status, year, and gap between year and hiring year. Standard errors are clustered at the teacher/hiring year level. R^2 values without school fixed effects range from 0.1 to 0.2. *** $p < .01$, ** $p < 0.05$, * $p < 0.10$.

[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.

[b]With the exception of Specifications 5 and 6, each coefficient is estimated in a separate regression.

Table 8. Predictors of Teacher Attrition from District

	District Attrition	
		School FE
(Spec. 1) 21-Point Score	N = 1,211 (618)[a]	N = 1,038 (554)
21-Point Score	-0.159*	-0.137
	(0.094)	(0.101)
(Spec. 2) 21-Point Components[b]	N = 851 (468)	N = 691 (400)
Experience	-0.063	-0.115
	(0.108)	(0.114)
Depth of Skills	-0.185*	-0.142
	(0.112)	(0.120)
Recommendations	-0.150	-0.073
	(0.121)	(0.127)
(Spec. 3) 60-Point Score	N = 1,266 (634)	N = 1,081 (567)
60-Point Score	-0.269***	-0.307***
	(0.104)	(0.114)
(Spec. 4) 60-Point Components	N = 1,266 (634)	N = 1,081 (567)
Certificate & Education	0.028	-0.009
	(0.112)	(0.122)
Training	-0.177*	-0.143
	(0.107)	(0.123)
Experience	-0.250**	-0.280**
	(0.104)	(0.119)
Classroom Management	-0.226**	-0.287**
	(0.097)	(0.113)
Flexibility	-0.234**	-0.252**
	(0.104)	(0.114)
Instructional Skills	-0.281***	-0.312**
	(0.105)	(0.122)
Interpersonal Skills	-0.336***	-0.377***
	(0.104)	(0.112)
Cultural Competency	-0.103	-0.129
	(0.104)	(0.112)
Preferred Qualifications	-0.231**	-0.271**
	(0.106)	(0.112)
Letters of Recommendation	-0.088	-0.148
	(0.116)	(0.129)
(Spec. 5) 21- and 60-Point Scores	N = 1,093 (561)	N = 930 (500)
21-Point Score	-0.216**	-0.182
	(0.105)	(0.111)
60-Point Score	-0.224**	-0.236*
	(0.112)	(0.122)
(Spec. 6) Factor Analysis	N = 766 (423)	N = 616 (360)
Factor 2 – 21-Point Score	-0.238**	-0.178
	(0.110)	(0.117)
Factor 1 – 60-Point Score	-0.284***	-0.343***
	(0.110)	(0.129)

Notes: Each of the specifications includes controls for gender, ethnicity, school size, school percentages for students eligible for free or reduced lunch, school percentages for under-represented minorities, and indicators for school level, Title I status, year, and gap between year and hiring year. Standard errors are clustered at the teacher level. Pseudo-R^2 values without school fixed effects are about 0.1. Sample sizes are smaller in school fixed effects models because some schools predict attrition perfectly.*** $p < .01$, ** $p < 0.05$, * $p < 0.10$

[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.

[b]With the exception of Specifications 5 and 6, each coefficient is estimated in a separate regression.

Table 9. Generalized First Stage Predicting Being Hired for Heckman Selection

	Hired	Given 60-Point Screen	Placebo (Hired Elsewhere)
21-Pt Screen	0.264***	0.425***	0.086**
	(0.023)	(0.019)	(0.038)
Excluded Variables:			
Error in Teacher's Favor	0.489***	0.661***	-0.131
	(0.063)	(0.045)	(0.108)
21-Pt Screen Competition	-0.416***	-0.464***	-0.030
	(0.067)	(0.040)	(0.057)
Observations	41,869 (3,939) [a]	41,869 (3,939)	41,369 (3,844)
F(Excluded Variables)	103.58***	366.61***	1.79

Models are estimated using probit with an unreported constant term. No additional controls are included. *** $p < .01$, ** $p < 0.05$, * $p < 0.10$

[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.

Table 10. The Effect of Screening Scores on Outcomes, With and Without Selection Correction

	Math		Reading	
VARIABLES	(1)	(2)	(3)	(4)
21-Pt Screen	0.036	0.035	0.028	0.026
	(0.045)	(0.046)	(0.029)	(0.029)
60-Pt Screen	0.050	0.048	0.018	0.007
	(0.054)	(0.055)	(0.056)	(0.056)
Mills Ratio (λ)		-0.076		-0.243
		(0.146)		(0.150)
Observations	73 (59) [a]		69 (56)	
R-Squared	0.178		0.122	
Overidentification p-value		0.901		0.835
	Absences		Monday/Friday Absences	
21-Pt Screen	1.167*	1.131	0.143	0.089
	(0.684)	(0.709)	(0.360)	(0.379)
60-Pt Screen	-0.119	-0.180	0.146	0.057
	(0.704)	(0.708)	(0.425)	(0.425)
Mills Ratio (λ)		-0.808		-1.178
		(1.743)		(0.994)
Observations	140 (106)		140 (106)	
R-Squared	0.355		0.270	
Overidentification p-value		0.704		0.843
	1-Year District Attrition		3-Year District Attrition	
21-Pt Screen	-0.039	-0.043	-0.083	-0.080
	(0.048)	(0.047)	(0.053)	(0.051)
60-Pt Screen	-0.079	-0.078	-0.110	-0.117
	(0.103)	(0.100)	(0.108)	(0.108)
Mills Ratio (λ)		-0.084		0.039
		(0.103)		(0.114)
Observations	195 (190)		135 (131)	
R-Squared	0.625		0.710	
Overidentification p-value		0.917		0.812

Estimates are produced using Specification 3 as presented in **Tables 6-8**, except that the sample is limited to those hired into Spokane in the sampling window, a linear probability model is used for attrition, and the selection correction as generated in **Table 9** is included in models (2) and (4). *** $p < .01$, ** $p < 0.05$, * $p < 0.10$

[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.

Table 11: Component Weights Which Maximize Correlation with Outcomes

		Math Value-Added	Reading Value-Added	Absences	1-Yr District Attrition
60-Point Rubric Component Weights	Certificate & Education	0.109	0.139	0.286	0.171
	Training	0.102	0.127	0.063	0.053
	Experience	0	0	0	0.069
	Classroom Management	0.512	0.158	0	0.068
	Flexibility	0	0.117	0.034	0.022
	Instructional Skills	0.043	0	0.031	0
	Interpersonal Skills	0.075	0.025	0	0.236
	Cultural Competency	0	0	0.141	0
	Preferred Qualifications	0.009	0.354	0.060	0.178
	Letters of Recommendation	0.150	0.080	0.385	0.203
Coefficient of *Weighted* 60-Pt Screening Score		0.144**	0.051	-0.405	-0.535**
		(0.042)	(0.036)	(0.556)	(0.222)
Standard (equally)-weighted Model Coefficients		0.074**	0.033	-0.054	-0.269***
		(0.028)	(0.025)	(0.503)	(0.104)

Note: Models include 21-point pre-screening rubric ratings and other controls identified in **Tables 6-8**. Comparison coefficients on non-weighted 21-point screening scores are limited to those for whom 21-point component scores are observed. *** $p < .01$, ** $p < 0.05$, * $p < 0.10$

Figure 1. Spokane Public Schools Hiring Process

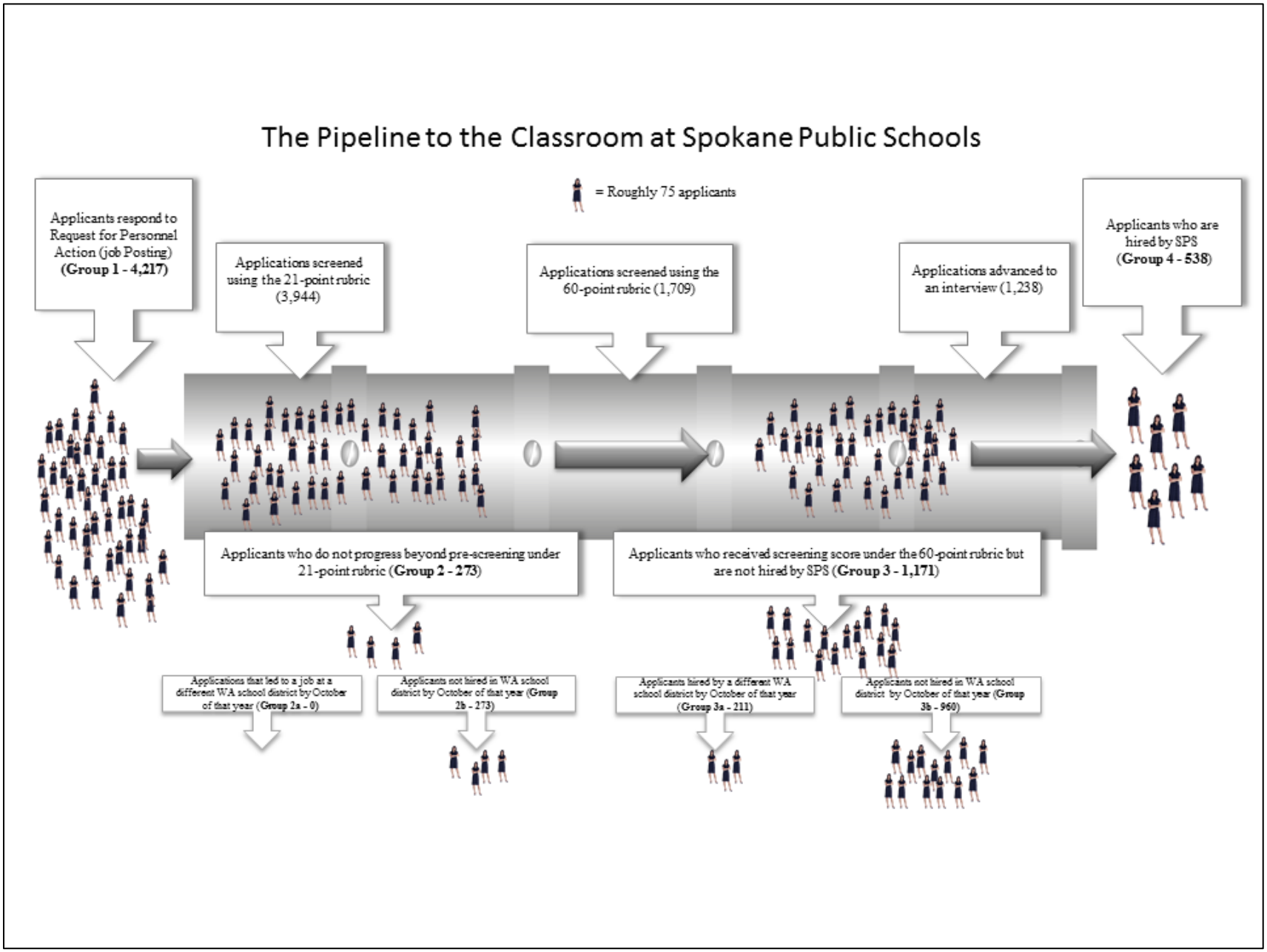
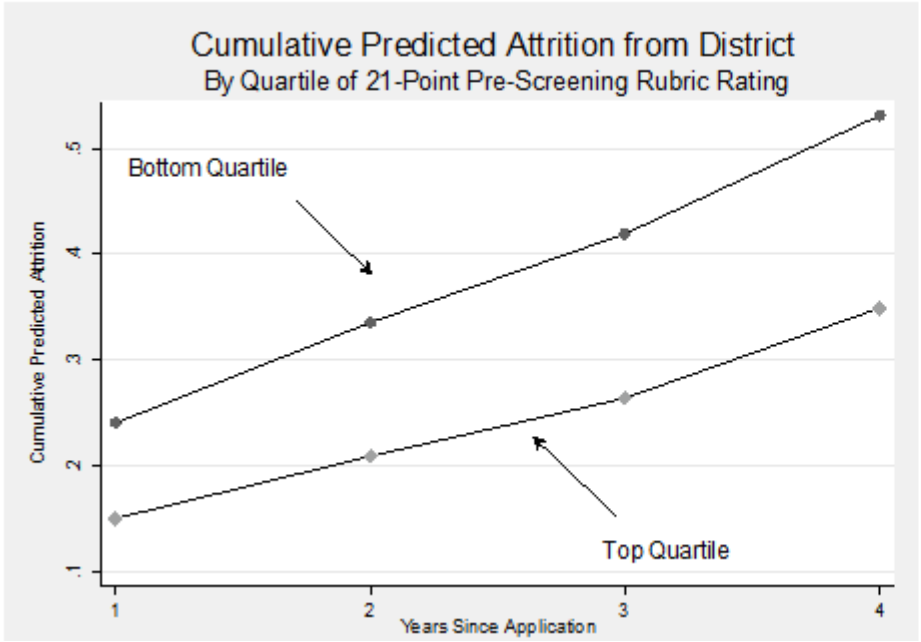
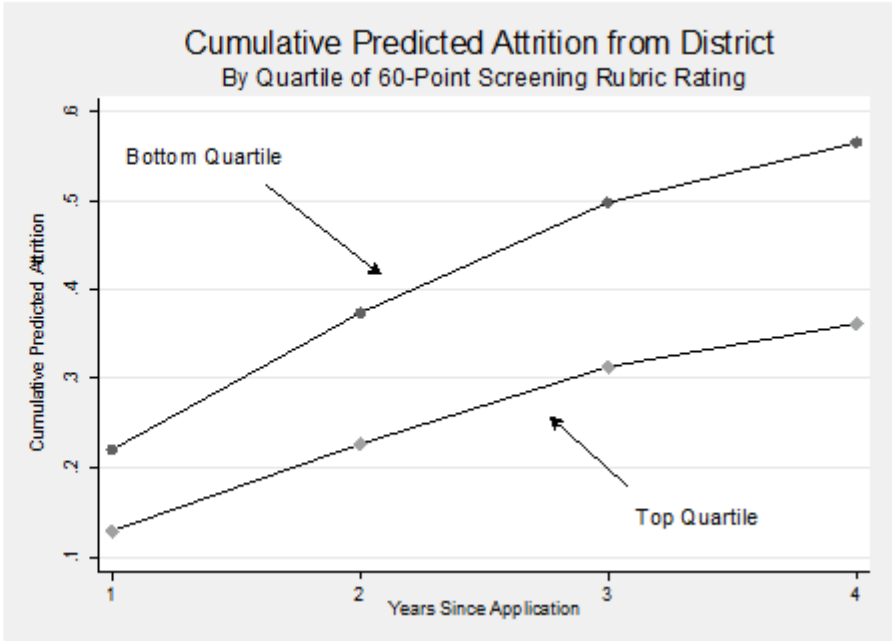


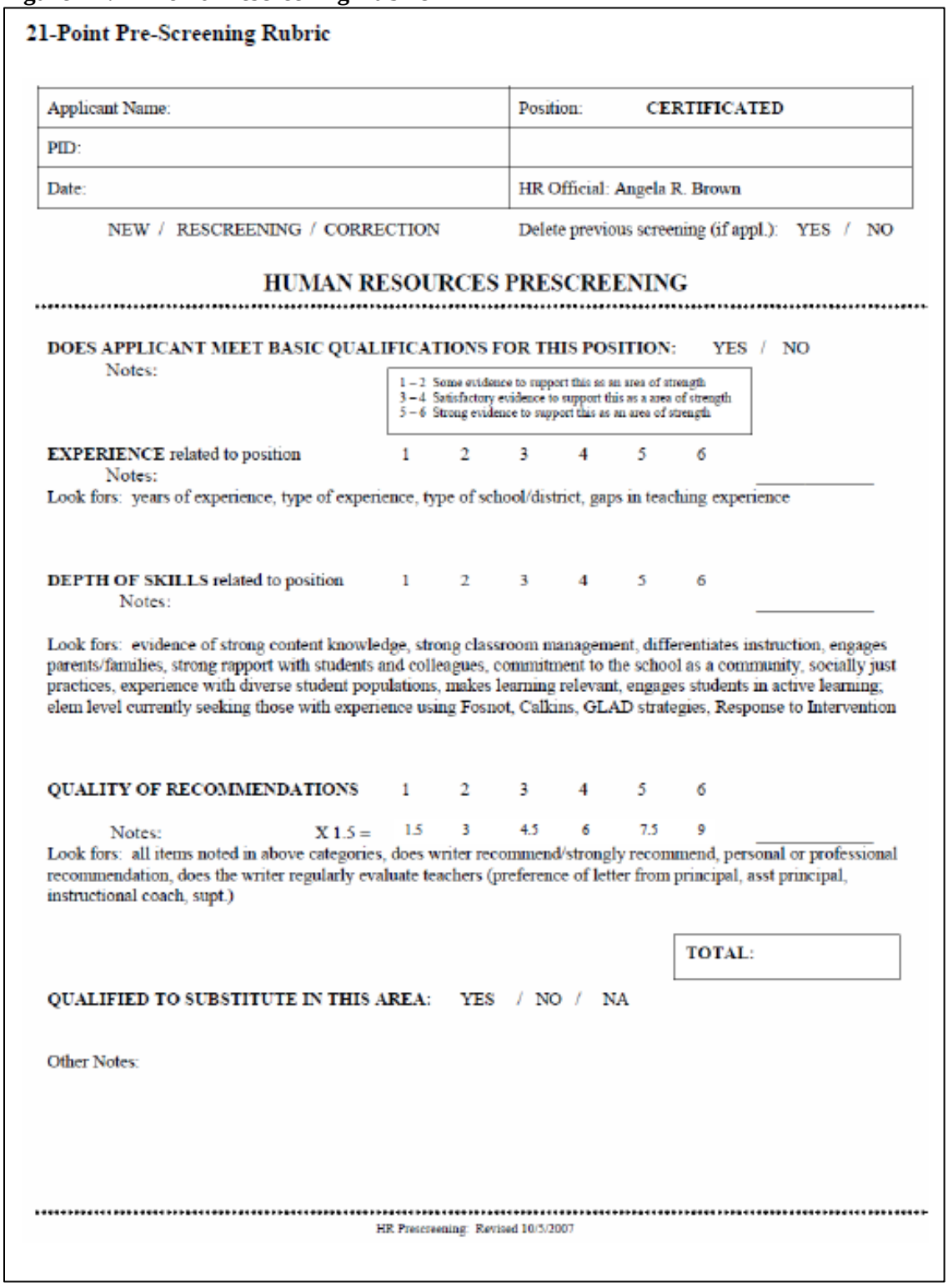
Figure 2. Predicted Probability of Attrition by Bottom and Top Quartile Scores on Screening Rubrics





Appendix A—Screening Rubrics and Generation of Applicant Data

Figure A1. 21-Point Prescreening Rubric



21-Point Pre-Screening Rubric

Applicant Name:	Position: **CERTIFICATED**
PID:	
Date:	HR Official: Angela R. Brown

NEW / RESCREENING / CORRECTION Delete previous screening (if appl.): YES / NO

HUMAN RESOURCES PRESCREENING

DOES APPLICANT MEET BASIC QUALIFICATIONS FOR THIS POSITION: YES / NO

Notes:

1 – 2 Some evidence to support this as an area of strength
3 – 4 Satisfactory evidence to support this as a area of strength
5 – 6 Strong evidence to support this as an area of strength

EXPERIENCE related to position 1 2 3 4 5 6

Notes:

Look fors: years of experience, type of experience, type of school/district, gaps in teaching experience

DEPTH OF SKILLS related to position 1 2 3 4 5 6

Notes:

Look fors: evidence of strong content knowledge, strong classroom management, differentiates instruction, engages parents/families, strong rapport with students and colleagues, commitment to the school as a community, socially just practices, experience with diverse student populations, makes learning relevant, engages students in active learning, elem level currently seeking those with experience using Fosnot, Calkins, GLAD strategies, Response to Intervention

QUALITY OF RECOMMENDATIONS 1 2 3 4 5 6

Notes: X 1.5 = 1.5 3 4.5 6 7.5 9

Look fors: all items noted in above categories, does writer recommend/strongly recommend, personal or professional recommendation, does the writer regularly evaluate teachers (preference of letter from principal, asst principal, instructional coach, supt.)

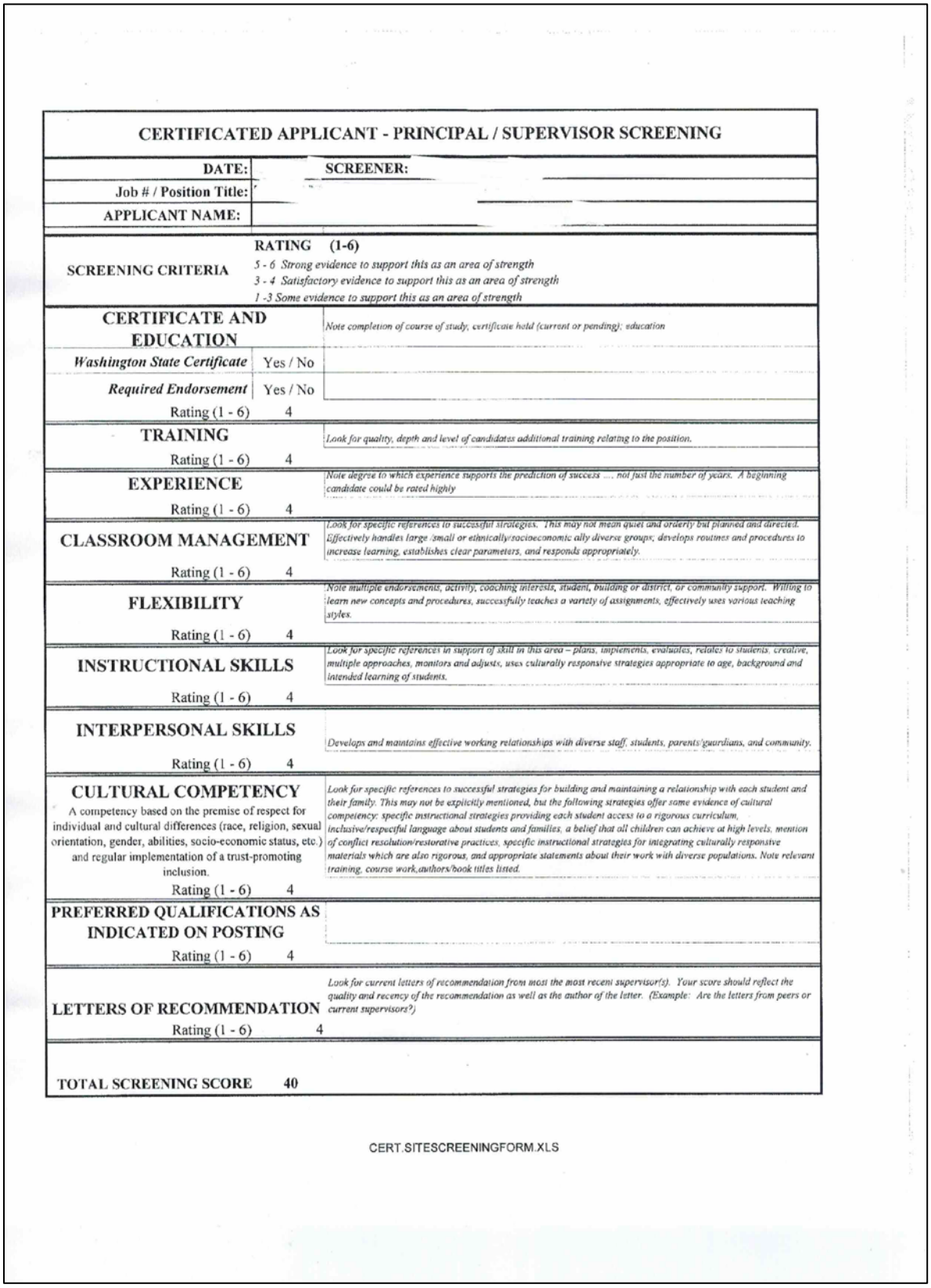
TOTAL:

QUALIFIED TO SUBSTITUTE IN THIS AREA: YES / NO / NA

Other Notes:

HR Prescreening: Revised 10/5/2007

Figure A2. 60-Point Screening Rubric



CERTIFICATED APPLICANT - PRINCIPAL / SUPERVISOR SCREENING

DATE:	**SCREENER:**
Job # / Position Title:	
APPLICANT NAME:	

SCREENING CRITERIA	**RATING (1-6)** *5 - 6 Strong evidence to support this as an area of strength* *3 - 4 Satisfactory evidence to support this as an area of strength* *1 -3 Some evidence to support this as an area of strength*
CERTIFICATE AND EDUCATION	*Note completion of course of study, certificate held (current or pending); education*
Washington State Certificate Yes / No	
Required Endorsement Yes / No	
Rating (1 - 6) 4	
TRAINING	*Look for quality, depth and level of candidates additional training relating to the position.*
Rating (1 - 6) 4	
EXPERIENCE	*Note degree to which experience supports the prediction of success .... not just the number of years. A beginning candidate could be rated highly*
Rating (1 - 6) 4	
CLASSROOM MANAGEMENT	*Look for specific references to successful strategies. This may not mean quiet and orderly but planned and directed. Effectively handles large /small or ethnically/socioeconomic ally diverse groups; develops routines and procedures to increase learning, establishes clear parameters, and responds appropriately.*
Rating (1 - 6) 4	
FLEXIBILITY	*Note multiple endorsements, activity, coaching interests, student, building or district, or community support. Willing to learn new concepts and procedures, successfully teaches a variety of assignments, effectively uses various teaching styles.*
Rating (1 - 6) 4	
INSTRUCTIONAL SKILLS	*Look for specific references in support of skill in this area – plans, implements, evaluates, relates to students, creative, multiple approaches, monitors and adjusts, uses culturally responsive strategies appropriate to age, background and intended learning of students.*
Rating (1 - 6) 4	
INTERPERSONAL SKILLS	*Develops and maintains effective working relationships with diverse staff, students, parents/guardians, and community.*
Rating (1 - 6) 4	
CULTURAL COMPETENCY A competency based on the premise of respect for individual and cultural differences (race, religion, sexual orientation, gender, abilities, socio-economic status, etc.) and regular implementation of a trust-promoting inclusion.	*Look for specific references to successful strategies for building and maintaining a relationship with each student and their family. This may not be explicitly mentioned, but the following strategies offer some evidence of cultural competency: specific instructional strategies providing each student access to a rigorous curriculum, inclusive/respectful language about students and families, a belief that all children can achieve at high levels, mention of conflict resolution/restorative practices, specific instructional strategies for integrating culturally responsive materials which are also rigorous, and appropriate statements about their work with diverse populations. Note relevant training, course work,authors/book titles listed.*
Rating (1 - 6) 4	
PREFERRED QUALIFICATIONS AS INDICATED ON POSTING	
Rating (1 - 6) 4	
LETTERS OF RECOMMENDATION	*Look for current letters of recommendation from most the most recent supervisor(s). Your score should reflect the quality and recency of the recommendation as well as the author of the letter. (Example: Are the letters from peers or current supervisors?)*
Rating (1 - 6) 4	
TOTAL SCREENING SCORE 40	

CERT.SITESCREENINGFORM.XLS

Job Postings

The hiring process for both new and replacement positions begins at the school level with a Request for Personnel Action. The request requires approval from Budget Authority, Accounting, and Human Resources. Postings are reviewed for the days, hours, and skills required, whether the position is a continuing contract or 1-year-only (OYO), the full-time-equivalent (FTE) level, and location. After a final review by the principal or position supervisor, the position is posted on the Spokane Public Schools (SPS) website (http://www.spokaneschools.org/page/2301). The posting provides detailed information about the position, including location, FTE level, contract type, anticipated start date, certification requirements, and preferred endorsements. The posting also provides detailed lists of responsibilities and qualifications, as well as application instructions and terms of employment. Jobs are posted for a minimum of 5 days, and hard-to-fill positions may post for as long as 10 days.

The study sample is restricted to certificated classroom teaching positions. A position is "certificated" if it requires personnel who hold a valid teaching credential. We characterize a "classroom teaching" position as one in which the teacher spends the majority of his or her time instructing students. Examples of certificated positions that we exclude from the study are nurse, counselor, and instructional coach. Finally, the sample is restricted to jobs for which we were able to obtain at least one 60-point screening score.[45] In total, the job set consists of 521 job postings. The types of job postings included and excluded from the analysis are listed in **Table A1**.

Table A1. Job Posting Categories

Included	Frequency	Excluded	Frequency
Engineering	14	Alternative Education	15
English Language Dev	12	Alternative Program	9
English Teacher	21	Career/College Prep Teacher	19
Foreign Lang Teacher	28	Core Support	5
Elementary Teacher	150	Counselor	9
Health/Fitness	11	Facilitator	12
Kindergarten Teacher	66	ICAN[a]	9
Math	42	Instructional Coach	9
Music/Arts	23	Intervention	75
Reading Teacher	6	Library Media Specialist	26
Science	31	Nurse	2
Social Studies Teacher	30	Odyssey Program	4
Special Education	63	Behaviorally Impaired	11
Teacher Or Substitute	13	Subject Coach	17
Vocational	11	Therapist/Psychologist	11
		Virtual Learning	3

[a]ICAN is the Individual Credit Advancement Now program and consists of specially designed web-based virtual learning courses.

[45] Positions for which a senior internal transfer was hired often do not have a 60-point screening score because the two most senior internal transfer applicants with the required qualifications are automatically granted interviews.

Applicants and Applications

Job applications are submitted and processed using SPS's online application system, which utilizes applicant management software from WinOcular (see www.winocular.com). In order to apply for any job, an applicant must create a user ID and password, and generate an applicant profile. Here, applicants are guided through the submission of information about their qualifications, as outlined in **Table A2.**

Table A2. Applicant Profile Fields

Employment Preferences	• Applicant can identify up to five areas of preference in terms of school level (e.g., middle school) and subject area (e.g., math)
Education	• College, degree, major, and dates attended (reporting of GPA and submission of transcripts is optional) • Additional trainings, classes, or workshops
Qualifications	• Certificates: identification number, type, and issue date • Endorsements: grade level and subject area
Experience	• List employers, dates, and reason for leaving o Credentialed experience o Student teaching/practicum/internship o General work experience
References	• Position and contact information • At least three letters of recommendation are required
Narrative Statements (2,000 character limit):	• "Describe how you will support a safe and academically rigorous learning environment for a wide variety of diverse student populations (including race, religion, sexual orientation, gender, abilities, socioeconomic status, etc.)" • "State briefly how and what you can contribute to Spokane Public Schools. Also include any other pertinent information that could assist in the evaluation of your application."
Supporting Documentation	• Cover letter • Resume • Copies of certificates and endorsements

Applicants can create multiple "applications" targeting different types of jobs. For example, an applicant applying to both certified teaching positions and classified positions will generally have separate applications for those job types, with different cover letters and narrative statements. Different applications can also be created for different types of certified positions (e.g., elementary and middle school, or classroom teacher and counselor). To apply for a specific job, an applicant selects the desired position and associates an application with it. Particularly for similar types of job postings (elementary teacher positions at two different schools, for example), the time required to apply for an additional position can be quite small, and as we explain below, many applicants submit a large volume of applications.

When the job posting is closed, an applicant list is created and all applications are checked for completion. If the application is not completed, an email is sent to the applicant stating what is required to finish the application. Applications are also checked for background disclosure statements and to determine whether an applicant is a voluntary internal transfer applicant.

Before external applicants are considered, the applicant list is reviewed for internal transfer applicants. Interviews are automatically granted to the two most senior internal applicants, properly certified employees requesting a transfer who have completed the application process. If an internal transfer candidate is not hired, HR must receive documentation of reasons for nonselection and an HR manager may request further information from the principal or supervisor. Once the nonselection notification is made, the external applicant list can be sent to the hiring team for screening. A school may conduct interviews with these other applicants, but a job offer cannot be made until 5 days after the nonselection notification.

As a first step in constructing the applicant data set, we identify every applicant who applied to one or more jobs in the set of 521 job postings, defined above. The profile for each of these applicants (2,669 unique individuals and 4,217 unique applicant-year combinations) is extracted from WinOcular as a data set with one observation per applicant. A second step is to extract each applicant-job combination with one observation per *application*. While the typical applicant submits multiple applications during a hiring year (mean = 10.1, median = 5) we observe only one applicant profile per applicant. This is in contrast to the screening data, which are year-specific (in the case of the 21-point prescreening scores) and job-specific (in the case of the 60-point screening scores).

Applicant Screening Instruments

Data for the 21-point prescreening rubric are exported from the WinOcular database and include the total score entered by the screener and the scores (1–6 points) for each of the three criteria—measures of teachers' value-added contributions to student learning, teacher absence behavior, and attrition rates. Most applicants have only one 21-point score in a hiring year. However, an applicant may have multiple scores if, for example, he or she applies for a different type of position from the current or previous one, or requests to be rescreened to reflect new experiences or skills. Where there are multiple scores, we associate each job application with the first screening score postdating the job application if that screening score was conducted within 30 days of the application. If there is not a screening score immediately postdating the application, we associate the job application with the most recent screening score that predates the application.[46]

As discussed in Section 3.3, each applicant who progresses beyond the initial screening is scored on a 60-point rubric. The screening scores associated with each job are kept as paper records in a job file folder at HR. The folder for each job file is labeled with the job posting number. For the certified positions we are interested in, these posting numbers are of the format *TE###-YY. TE* indicates that it is a certificated position, *###* is a three-digit number that restarts at "001" each hiring year, and *YY*

[46] Among applicants in the study sample, 71% have one 21-point score, 21% have two scores, and 6% have three scores.

indicates the hiring year. For example, TE025-12 would be the 25th certified job positing in the hiring year starting in March 2012 and ending in February 2013. In this paper, we refer to this time period as the “2012 hiring year.” Records for the 2009 through 2012 hiring years are available to this project (older records are destroyed).

The 60-point rubrics with recorded scores for each applicant, as well as the interview request forms, were pulled from the job folders and scanned into PDF files (one file per job). The scanned job file data were manually entered into a spreadsheet including the following data fields: job posting number, applicant name (first and last), screener name (first and last), score on each of the rubric’s criteria, total score reported by the screener, and whether an interview was requested (yes/no). The screening score data are matched to the data derived from WinOcular by the applicant’s first and last name and the job posting number.

Appendix B—Supplemental Descriptive and Regression Tables

Table B1. Applications per Job Posting, by Job Type

	School Level and Hiring Year											
	Elementary				Middle School				High School			
	2009	2010	2011	2012	2009	2010	2011	2012	2009	2010	2011	2012
Engineering	-	-	-	-	-	-	-	8	10	6	7	-
English Teacher	34	31	47	28	-	-	25	69	35	33	31	52
Foreign Lang Teacher	-	-	-	-	-	11	7	6	7	7	5	5
Elementary Teacher	137	176	158	182	-	-	-	-	-	-	-	-
Health/Fitness	-	-	-	-	-	-	-	14	24	30	-	13
Kindergarten Teacher	112	143	131	129	-	-	-	-	-	-	-	-
Math	-	-	-	-	18	49	37	43	19	23	24	10
Music/Arts	-	-	-	-	3	11	9	12	10	11	-	7
Reading Teacher	69	69	-	-	-	-	-	67	-	-	-	-
Science	-	-	-	-	-	31	46	26	12	20	16	15
Social Studies Teacher	-	-	-	-	34	68	43	57	35	41	29	61
Special Education	37	44	28	36	36	44	-	38	28	33	39	32

Table B2: Factors Generated Using Factor Analysis

Factors of Screening Scores		Factor 1	Factor 2	
17-Point Components	Experience		0.338	
	Depth of Skills		0.790	
	Letters of Recommendation		0.723	
60-point rubric component weights	Certificate & Education	0.316		
	Training	0.595		
	Experience	0.679		
	Classroom Management	0.788		
	Flexibility	0.791		
	Instructional Skills	0.818		
	Interpersonal Skills	0.784		
	Cultural Competency	0.623		
	Preferred Qualifications	0.433		
	Letters of Recommendation	0.400		
Factors of Objective Criteria		**Factor 1**	**Factor 2**	**Factor 3**
Taught in SPS				
Known College			0.703	
Student Taught SPS			0.682	
WESTB Reading		0.656		
WESTB Math		0.547		
WESTB Writing		0.649		
College GPA				0.350
SAT Combined				0.391
Master's Degree				0.518

Promax rotation is used. Components with factor loadings below .3 are not shown.

Table B3. Primary Outcome Models Restricted & Unrestricted by SPS Employment

		Student Outcomes		Teacher Absences		Attrition
		Math (1)	Reading (2)	Total Absences (3)	Mon. & Fri. Absences (4)	District (6)
Restricted	21-Point Rubric Summative Score	0.016 (0.024)	0.029 (0.021)	0.486 (0.702)	0.238 (0.321)	-0.216** (0.105)
	60-Point Rubric Summative Score	0.048 (0.035)	0.003 (0.032)	-0.091 (0.523)	-0.010 (0.288)	-0.224** (0.112)
	Observations	132 (107) [a]	128 (104)	272 (205)		1,093 (561)
	R-squared	0.158	0.101	0.154	0.133	0.072
Unrestricted	**Inside Spokane**					
	21-Point Rubric Summative Score	0.033 (0.037)	0.034 (0.026)	1.290* (0.677)	0.330 (0.340)	-0.245* (0.142)
	60-Point Rubric Summative Score	0.035 (0.050)	-0.008 (0.041)	0.066 (0.581)	-0.061 (0.386)	-0.252* (0.143)
	Observations	94 (72)	92 (71)	178 (134)		780 (391)
	R-squared	0.141	0.152	0.198	0.158	0.072
	Outside of Spokane					
	21-Point Rubric Summative Score	0.008 (0.041)	0.011 (0.049)	-1.975 (1.634)	-0.676 (0.708)	-0.242 (0.176)
	60-Point Rubric Summative Score	0.124** (0.051)	-0.004 (0.050)	-0.283 (1.114)	0.254 (0.460)	-0.135 (0.209)
	Observations	38 (36)	36 (33)	94 (71)		313 (177)
	R-squared	0.536	0.250	0.267	0.266	0.133

Notes: All regressions displayed in this table are run with identical controls and predictor variables as the primary outcome models above. Subsample observations do not add up to the full sample due to several teachers who taught both inside and outside Spokane in the same year. **** p < .01, ** p < 0.05, * p < 0.10*.

[a] The number of clusters in each analysis is presented in parentheses next to the total number of observations.

Table B4. Primary Outcome Models Restricted and Unrestricted, by School Level

		Student Outcomes		Teacher Absences		Attrition
		Math (1)	Reading (2)	Total Absence (3)	Monday-Friday Absences (4)	District (6)
Restricted	21-Point Rubric Summative Score	0.016 (0.024)	0.029 (0.021)	0.486 (0.702)	0.238 (0.321)	-0.216** (0.105)
	60-point Rubric Summative Score	0.048 (0.035)	0.003 (0.032)	-0.091 (0.523)	-0.010 (0.288)	-0.224** (0.112)
	Observations	132 (107) [a]	128 (104)	272 (205)		
	R-squared	0.158	0.101	0.154	0.133	0.072
Unrestricted	**Elementary**					
	21-Point Rubric Summative Score	0.023 (0.032)	0.027 (0.025)	-0.210 (0.835)	0.086 (0.386)	-0.282* (0.157)
	60-Point Rubric Summative Score	0.013 (0.043)	-0.009 (0.034)	-0.517 (0.665)	-0.230 (0.321)	-0.143 (0.169)
	Observations	95 (76)		138 (102)		603 (302)
	R-squared	0.085	0.069	0.294	0.343	0.071
	Middle school					
	21-Point Rubric Summative Score	0.055 (0.043)	0.054 (0.036)	1.161 (0.855)	0.371 (0.392)	0.048 (0.398)
	60-Point Rubric Summative Score	0.092* (0.054)	-0.065 (0.057)	0.807 (1.378)	-0.041 (0.752)	-0.649* (0.333)
	Observations	37 (32)	34 (32)	57 (44)		168 (92)
	R-squared	0.544	0.424	0.416	0.313	0.247
	High school					
	21-Point Rubric Summative Score	-	-	1.300 (0.895)	0.339 (0.557)	-0.127 (0.189)
	60-Point Rubric Summative Score	-	-	0.118 (0.943)	0.423 (0.672)	-0.212 (0.239)
	Observations	-	-	76 (60)		286 (160)
	R-squared	-	-	0.159	0.157	0.193

Notes: All regressions displayed in this table are run with identical controls and predictor variables as the primary outcome models above. Subsample observation numbers do not add up to the full sample because some teachers taught at both levels in the same year, and in attrition subsample regressions because some observations were dropped due to perfect prediction in the subsample. *** $p < .01$, ** $p < 0.05$, * $p < 0.10$

[a] The number of clusters in each analysis is presented in parentheses next to the total number of observations.

Table B5. Predictors of Teacher Attrition from School and State

	School		State	
		School FE		School FE
(Spec. 1) 21-point score	N = 1,133 (613)[a]	N = 1,027 (611)	N = 1,266 (622)	N = 1,071 (562)
21-point score	-0.227*** (0.080)	-0.183** (0.087)	-0.169 (0.126)	-0.144 (0.148)
(Spec. 2) 21-point components	N = 800 (464)	N = 786 (455)	N = 879 (465)	N = 714 (403)
Experience	-0.125 (0.088)	-0.195** (0.097)	-0.013 (0.136)	-0.078 (0.144)
Depth of skills	-0.220** (0.098)	-0.149 (0.110)	-0.202 (0.159)	-0.179 (0.173)
Recommendations	-0.201** (0.093)	-0.123 (0.107)	-0.158 (0.153)	-0.092 (0.173)
(Spec. 3) 60-point score	N = 1,182 (627)	N = 1,164 (622)	N = 1,322 (639)	N = 1,113 (574)
60-point score	-0.276*** (0.091)	-0.331*** (0.097)	-0.308*** (0.119)	-0.391*** (0.133)
(Spec. 4) 60-point components [b]	N = 1,182 (627)	N = 1,164 (622)	N = 1,322 (639)	
Certificate & education	0.027 (0.096)	0.026 (0.101)	0.090 (0.130)	0.023 (0.147)
Training	-0.185** (0.091)	-0.189* (0.098)	-0.199 (0.127)	-0.266* (0.150)
Experience	-0.257*** (0.090)	-0.315*** (0.095)	-0.235* (0.122)	-0.337** (0.138)
Classroom management	-0.240*** (0.087)	-0.296*** (0.097)	-0.296*** (0.103)	-0.380*** (0.118)
Flexibility	-0.214** (0.092)	-0.245** (0.098)	-0.267** (0.116)	-0.301** (0.124)
Instructional skills	-0.223** (0.092)	-0.240** (0.099)	-0.371*** (0.117)	-0.433*** (0.133)
Interpersonal skills	-0.271*** (0.091)	-0.285*** (0.101)	-0.342*** (0.120)	-0.455*** (0.137)
Cultural competency	-0.145 (0.091)	-0.225** (0.100)	-0.218* (0.116)	-0.328** (0.134)
Preferred qualifications	-0.188** (0.092)	-0.236** (0.099)	-0.258** (0.119)	-0.288** (0.133)
Letters of recommendation	-0.218** (0.105)	-0.225** 0.026	0.061 (0.125)	0.036 (0.138)
(Spec. 5) 21- and 60-point scores	N = 1,023 (555)	N = 1,006 (550)	N = 1,146 (566)	N = 960 (508)
21-point score	-0.251*** (0.089)	-0.167* (0.095)	-0.212 (0.133)	-0.141 (0.159)
60-point score	-0.257** (0.102)	-0.322*** (0.111)	-0.279** (0.136)	-0.354** (0.150)
(Spec. 6) factor analysis	N = 723 (418)	N = 700 (406)	N = 795 (420)	N = 639 (363)
Factor 2 – 21-point score	-0.296*** (0.098)	-0.181* (0.106)	-0.194 (0.140)	-0.161 (0.163)
Factor 1 – 60-point score	-0.252*** (0.094)	-0.359*** (0.112)	-0.336** (0.131)	-0.470*** (0.154)

Notes: Each of the specifications includes controls for gender, ethnicity, school size, school percentages for students eligible for free or reduced lunch, school percentages for under-represented minorities, and indicators for school level, Title I status, year, and gap between year and hiring year. Standard errors are clustered at the teacher level. R^2 values are about 0.1. Sample sizes are lower in school fixed effect models because some school indicators predict success perfectly.
*** $p < .01$, ** $p < 0.05$, * $p < 0.10$

[a] The number of clusters in each analysis is presented in parentheses next to the total number of observations.

[b] With the exception of specifications 5 and 6, each coefficient is estimated in a separate regression.

Table B6. Predictive Validity of Screening Scores With Controls for Factors of Objective Criteria

	Math		Reading	
21-point score		-0.003		0.022
		(0.024)		(0.022)
60-point score		0.069**		0.020
		(0.034)		(0.031)
Factor 1 (WEST-B)	-0.038	-0.022	0.041	0.054
	(0.041)	(0.045)	(0.050)	(0.053)
Factor 2 (familiarity)	-0.147**	-0.161**	-0.072	-0.078
	(0.069)	(0.068)	(0.065)	(0.066)
Factor 3 (academics)	0.157**	0.159**	0.099	0.088
	(0.056)	(0.057)	(0.063)	(0.066)
Std. dev. of teacher effect	0.085	0.100	0.063	0.067
Observations	132 (107) [a]		128 (104)	
R^2	0.236	0.271	0.148	0.160
	Total absences		Monday–Friday absences	
21-point score		0.544		0.231
		(0.684)		(0.310)
60-point score		0.036		0.068
		(0.528)		(0.287)
Factor 1 (WEST-B)	-0.674	-0.651	-0.159	-0.137
	(0.954)	(0.952)	(0.480)	(0.476)
Factor 2 (familiarity)	1.127	1.310	0.524	0.668
	(1.137)	(1.108)	(0.623)	(0.618)
Factor 3 (academics)	0.888	0.836	0.945	0.924
	(1.291)	(1.266)	(0.676)	(0.668)
Std. Dev. of Teacher Effect	0.522	0.709	0.350	0.412
Observations	272 (205)		272 (205)	
R^2	0.158	0.162	0.078	0.079
	District attrition			
21-point score		-0.242**		
		(0.103)		
60-point score		-0.202*		
		(0.111)		
Factor 1 (WEST-B)	0.011	-0.004		
	(0.215)	(0.219)		
Factor 2 (familiarity)	-0.265	-0.324		
	(0.265)	(0.270)		
Factor 3 (academics)	-0.001	0.043		
	(0.258)	(0.259)		
Std. dev. of teacher effect	0.124	0.342		
Observations	1,093 (561)			
R^2	0.053	0.065		

Notes: All regressions displayed in this table are run with identical controls and predictor variables as the specification 6 models presented in **Tables 6-8**, except with the addition of the factors described in **Table B2**.
*** $p < .01$, ** $p < 0.05$, * $p < 0.10$
[a]The number of clusters in each analysis is presented in parentheses next to the total number of observations.