# Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools

Brian A. Jacob, University of Michigan

Jonah E. Rockoff, Columbia Business School

Eric S. Taylor, Harvard Graduate School of Education

Benjamin Lindy, Teach For America

Rachel Rosen, MRDC

August 2018*

Selecting more productive employees among a pool of job applicants can be a cost-effective means of improving organizational performance and may be particularly important in the public sector. We study the relationship among applicant characteristics, hiring outcomes, and job performance for teachers in the Washington DC Public Schools. Applicants' academic background (e.g., undergraduate GPA) is essentially uncorrelated with hiring. Screening measures (written assessments, interviews, and sample lessons) help applicants get jobs by placing them on a list of recommended candidates, but they are only weakly associated with the likelihood of being hired conditional on making the list. Yet both academic background and screening measures strongly predict teacher job performance, suggesting considerable scope for improving schools via the selection process.

JEL No. I2, J2, M51

* Jacob: bajacob@umich.edu, 735 South State Street, Ann Arbor, MI 48109. Rockoff: jonah.rockoff@columbia.edu, 3022 Broadway #603, New York, NY 10027. Taylor: eric_taylor@harvard.edu, Gutman Library 469, 6 Appian Way, Cambridge, MA 02138. Lindy: benjamin.lindy@gmail.com, 1110 Main St., Cincinnati, OH 45202. Rosen: rachel.rosen@mdrc.org, 19th Floor, 16 East 34 Street, New York, NY 10016.

We first thank the District of Columbia Public Schools, in particular Michael Gaskins, Anna Gregory, Brooke Miller, Jason Kamras, and Scott Thompson. Generous financial support was provided by the Smith Richardson Foundation. We received helpful comments and suggestions from seminar participants at Brown, Chicago, Clemson, Cornell, Delaware, Johns Hopkins, Kentucky, LSU, New York Fed, NYU, Northwestern University, Paris School of Economics, Princeton, Stanford, Texas A&M, UC Santa Barbara, Universidad Nacional La Plata, Universidad San Andres, University of Warwick, APPAM, and AEFP.

The authors of this publication were consultants to the District of Columbia Public Schools. The terms of this relationship and this publication have been reviewed and found to be in accordance with the DCPS policy on objectivity in research by the Office of Talent and Culture and by the Office of Instructional Practice District of Columbia Public Schools.

*"The best means of improving a school system is to improve its teachers. One of the most effective means of improving the teacher corps is by wise selection."*

Ervin Eugene Lewis, Superintendent of Schools, Flint, Michigan, 1925

The importance of employee selection is widely recognized by practitioners and researchers alike. While a large literature in psychology has explored the power of applicant screening to identify successful employees (see, for example, McDaniel et al. 1994), economists have paid far less attention to this issue. As noted by Oyer and Schaeffer (2011), in contrast to economists' work on understanding employee incentives, "the literature has been less successful at explaining how firms can find the right employees in the first place." This has recently begun to change, with papers on the role of personal referrals (Schmutte 2015, Burks et al. 2015, Brown et al. 2016), placement agencies (Stanton and Thomas 2016), objective screening technologies (Hoffman et al. 2015), and recruitment messaging (Ashraf et al. 2016) in the hiring process.

There are several reasons why employee selection is particularly important in the field of public education. First, there is substantial variance in teacher effectiveness, and good teachers have positive impacts on long-term student outcomes (see Koedel et al. 2015 for a review). Second, there are substantial financial and political costs to removing teachers who perform poorly on the job.[1] Third, in most parts of the country and in most subject areas there is an abundance of *potential* teachers (see Greenberg et al. 2013), yet most prior research suggests that school systems are not very good at selecting the individuals most likely to become successful teachers (see Ballou 1996, Kane and Staiger 2005, Harris et al. 2014, Hinrichs 2014).

[1] Barnes et al. (2007) estimate that turnover costs districts roughly $10,000 per teacher. Staiger and Rockoff (2010) illustrate the academic cost of exposing students to newly hired but ineffective teachers, and Rothstein (2015) highlights the cost of compensating teachers for increasing the risk of job separation. In addition, collection of performance measures on teachers (e.g. standardized student testing, classroom observation, portfolios of student work) requires significant public resources and often entails difficult labor negotiations (e.g., Baker and Santora 2013), while schools and school districts have wide freedom in deciding what information to collect from applicants as part of the hiring process. Issues over teacher removal have also been the subject of major lawsuits (Treu 2014).

Only one study, in addition to ours, provides solid evidence on whether pre-employment measures from an actual job application can predict successful teaching.[2] Goldhaber et al. (2017) examine teacher applicants in Spokane Washington, where applications contain background information (education, experience, licensure), recommendation letters, and narrative statements. District personnel do an initial screen, then school-specific personnel fill out more detailed evaluations on candidates of interest. They find that teachers with higher rated applications (at either stage) have lower attrition rates, and they find a significant positive relationship between the school-based evaluation of teacher candidates and subsequent value-added in math (but not English).[3]

While these results are encouraging, the importance of the issue merits much further study. To that end, we use uncommonly detailed data on applications, job offers, employment, and performance of teachers in the Washington, DC Public Schools (hereafter DCPS) to make several contributions to the literature on employee selection in public schools.

First, we present an analysis of the hiring process. Previous work on labor demand for teachers has been limited, in contrast to the large literature on teacher labor supply, where it is often assumed (sometimes implicitly) that employment outcomes stem from teachers' choices, not those of school or district administration. We provide evidence on the extent to which schools use the information collected centrally by the district, and investigate whether

[2] Other studies that do speak to the issue of teacher selection suffer from important shortcomings stemming from the use of measures collected in low-stakes research surveys (Rockoff et al. 2011) or administrative data unavailable to schools and school districts (Boyd et al. 2008). Moreover, these studies only examine data on teachers who are already hired, rather than data on a pool of applicants, preventing them from addressing issues of selection. There is also a large literature outside economics that has studied the teacher hiring process, but this work is often qualitative in nature or relies on small samples with limited measures of teacher performance (see Appendix B for a more comprehensive discussion of this work).

[3] The estimated relationship between district screening scores and math value added is considerably smaller than the estimate for math and school-based scores, and is not statistically significant. Note that the authors also address potential biases from selective hiring by relying on arithmetic errors made in computing applicant scores.

applicants' characteristics are related to the types of schools where they receive job offers and are hired. Using data on both job offers and hiring outcomes provides additional support for the idea that our analysis is capturing demand-side factors.

Second, our analysis focuses on a broader measure of teacher performance than previous work. Our performance metric is based largely on classroom observations of teacher instruction and interaction with students, although it also incorporates a variety of other inputs such as supervisor ratings, student work, and (when available) value-added to students' standardized test scores. We also examine performance measured by test-score value-added alone, but our statistical power is limited.

The prevalence and high-stakes use of observation-based measures alone make them an important subject of inquiry. Over the past decade, new rubric-based classroom observation evaluations have been introduced by nearly all states (at least 46) and the nation's 25 largest districts (Steinberg and Donaldson 2016). These new observations are structured and scored using detailed rubrics, and generate considerable variation.[4]

Beyond their growing use in educational policy, classroom observation scores can capture important variation in teacher job performance, making them an informative measure of teacher quality. While value-added can be calculated for only about one-third of teachers in certain grades and subjects, all teachers can be measured with observations, and there is growing evidence that classroom observations do predict student achievement gains. Several studies report positive correlations (about 0.2) between observation scores and student test scores or

[4] Weisberg et al. (2009) provide evidence that "older generation" evaluations based on classroom observation that were more holistic often offer limited variation, with large fractions of teachers earning top scores. While final scalar evaluation scores (in DC and elsewhere) are also somewhat coarse, the raw data we use are collected in multiple observations, by multiple observers, scoring multiple practices, and provide meaningful variation.

teacher value-added (see, for example, Milanowski 2004, Kane et al. 2011, Kane and Staiger 2011, and Grossman et al. 2014) and we find similar correlations in our DCPS data.[5] Recent studies that randomly assign teachers to classes can rule out explanations due to within-school student sorting (see, for example, Kane et al. 2013, Garret and Steinberg 2015, Araujo et al. 2016, Bacher-Hicks et al. 2017).[6]

A teacher's job includes responsibilities beyond those reflected in test scores, and classroom observation scores also partly reflect non-test score student outcomes. Blazar and Kraft (2017) show that observation measures predict students' self-assessment of their math ability, happiness in class, and behavior in class. Similarly, Araujo et al. (2016) find observation scores predict executive functioning skills among kindergarten students.[7] In this sense our observation-based outcome is broader in the scope of teaching job responsibilities measured, although it is not a direct measure of student learning.

In our study, we examine the relationship between the teacher performance measures in DCPS and applicant characteristics. These include "traditional" measures of applicant quality (e.g., SAT score, college GPA) as well as measures based on candidates' writing, interviews, and auditions. To address potential bias from selection into hire, we exploit idiosyncratic features of the DCPS hiring process that create discontinuities between applicant scores and hiring/job

[5] The correlation between a teacher's value-added score in year $t$ and her classroom observation score in year $t$ is 0.27. The correlation between value-added in $t$ and mean observation score in years other than $t$ is 0.25.
[6] As further evidence against sorting bias, Bacher-Hicks et al. (2017) find that a teacher's observation score when students are assigned naturally is an unbiased predictor of the teacher's score when students are assigned randomly. White (2018) finds a similar result, as does Kane et al. (2013). Related evidence comes from Cantrell et al. (2008), who experimentally test the link between test scores and teachers' ratings by the National Board for Professional Teaching Standards; these ratings are based on a portfolio of teacher work including recorded lessons and tests of content knowledge. Additionally, there is some evidence that the process of being evaluated with classroom observations can itself improve teacher value-added to student test scores (Taylor and Tyler 2012).
[7] Other work has shown that teachers have effects on various other non-cognitive skills (Jackson 2016, Petek and Pope 2017). Given the evidence cited above, it seems likely that observation-based measures will predict these outcomes as well.

offers, and test the robustness of our findings to non-random sorting of teachers using specifications with school fixed effects.

Several interesting findings emerge. First, the district's less-traditional screening measures are strong predictors of teacher performance. Second, we find that several academic background characteristics (e.g., undergraduate GPA) also strongly predict subsequent teacher performance. Pooling all of these measures to create an index of predicted performance, we find the actual performance of "top quartile" hires is more than two-thirds of a standard deviation ($0.71\sigma$) higher than those from the bottom quartile.[8]

The predictive power we find for applicants' academic background may seem at odds with the stylized fact that observable characteristics other than teaching experience typically do not predict teacher effectiveness (Staiger and Rockoff 2010). This is indeed a common finding for many studies with access to observables typical of administrative data, like graduate degrees and certifications. However, several studies find predictive power for non-administrative observables such as test scores and other measures of teachers' academic achievement and cognitive ability (Clotfelter et al. 2007, 2010, Boyd et al. 2008, Rockoff et al. 2011, Taylor 2018). Ours is the first paper to link such non-administrative achievement measures to a broad, observation-based performance measure.

[8] To underscore the large magnitude of this finding, one can compare it to the average on-the-job improvement exhibited by DCPS teachers over their first three years working at DCPS, a period in teachers' careers when performance has been consistently shown to improve rapidly (e.g., Rice 2013, Ost 2014, Papay and Kraft 2015). Among the new teachers in our sample who remain in DCPS for three years, their average three-year growth in performance is 0.37 standard deviations, roughly half of the difference in performance between top- and bottom-quartile applicants entering DCPS.

Third, we find that predicted performance is only weakly, if at all, associated with an applicant's likelihood of receiving a job offer, being hired, or remaining as a teacher in DCPS. This is consistent with school principals not systematically hiring the most effective applicants and suggests considerable scope for improving teacher quality through the selection process.[9] However, since we do not observe if top candidates not hired by DCPS end up teaching elsewhere, our evidence cannot speak directly to selection at levels beyond the school district.

The paper proceeds as follows. In Section 2 we describe teacher application, hiring, and evaluation processes in DCPS. In Section 3 we describe our empirical strategy, and we present results on hiring, sorting, job performance, and attrition in Section 4. Section 5 concludes.

## 2. Application, Hiring, and Performance Evaluation in DCPS

Hiring in DCPS is largely decentralized, with school principals making decisions on whom to hire. In 2009, DCPS created TeachDC, a multi-stage, centralized application process, with the goal of helping principals connect to a set of "recommended" candidates and reduce principals' costs of screening out less desirable applicants. Our analysis focuses on applicants to TeachDC, but teachers can also apply directly for jobs at local schools, which may be common for those with pre-existing connections to school staff, or through special programs like Teach for America. We examine hiring from 2011 to 2013, during which roughly one-half of new hires applied to TeachDC and one-third of new hires came from those placed on the TeachDC recommended list. Although we lack data on non-TeachDC applicants, we find very little difference in the demographics of new hires who did and did not apply to TeachDC (Appendix Table 1).

[9] As mentioned above, this is also consistent with prior work (Ballou 1996; Kane and Staiger 2005; Hinrichs 2014), although Boyd et al. (2011) find evidence that principals may identify effective teachers among transfer applicants.

*2.1 The TeachDC Application Process*

Each year, from about February to July, candidates submit applications to TeachDC. The online application system first collects background information such as applicants' education history, employment experience, and eligibility for licensure.[10] Following collection of this preliminary information, district officials review applications in several stages, and each stage is scored by district personnel according to a standardized rubric. Applicants who pass a specified performance threshold proceed to the next stage. Appendix C has more details on the TeachDC process; here we provide the information most relevant to our analysis. Also, in 2011 DCPS piloted additional measures during the first stage of the application which were not used in the screening process. Appendix C discusses these measures in detail and shows estimates of how these predict teacher hiring and performance.[11]

After providing background information, applicants choose the subject and grade level combination to which they wish to apply and take a subject-and-level-specific written assessment of their pedagogical content knowledge (PCK) and knowledge of instructional practices.[12] Applicants who pass this stage are invited for an interview and, if they pass another cutoff score, are invited to a teaching "audition"; these were done in-person for some applicants in 2011 but, after facing logistical hurdles, TeachDC switched to having applicants submit video of themselves teaching in their subject area (e.g., from a student teaching experience or a prior job).

[10] Applicants who do not already hold a DC license and whose credentials make them ineligible to obtain one prior to the start of the school year are not allowed to proceed further, and we do not analyze these ineligible applications.
[11] There are two takeaways from this analysis. First, questions related to personality (conscientiousness, grit, neuroticism, etc.) do not yield useful data in a hiring setting because applicants simply provide the answers they think the employer wants to hear. Second, scores on a commercial teacher selection product (Haberman Star Teacher Pre-Screener) – to which the right answers are not obvious – have significant power to predict performance and are only weakly related to hiring, echoing our results for academics and screening scores.
[12] There were 28 subject and grade level combinations from which to choose, such as "Elementary (grades 1-5)", "Music (elementary)", "Middle school science (grades 6-8)", "High school science: Biology", and "Physical education / health".

Applicants who pass all stages are included in a recommended pool that is made available to principals online.[13] During the years we study, recommended applicants were listed in alphabetical order, with links to their resumes. Principals could filter the database by subject area and navigate through the online database to find out further information on how the applicants scored in the TeachDC process. While we know that DCPS principals were provided with an introduction to the database during a regular meeting of school administrators, the district did not track whether principals used the database, nor whether they proceeded beyond the list of candidates to view applicants' scores from the different hiring stages.

As mentioned above, the ultimate hiring decision is quite decentralized. Principals can use information from the TeachDC online database but they might also hire through personal recommendations from school staff, responses to independent job postings, or other recruitment channels. While we lack information on schools' hiring processes, existing evidence (e.g., Liu and Kardos 2002, Balter and Duncombe 2005) suggests principals identify promising candidates, ask them to present work samples and/or sit for interviews, and then decide on jobs offers. Nevertheless, nearly all schools (97.6%) hired at least one TeachDC applicant during the three years we study.

Figure 1 plots, by year, the probability of receiving a job offer and being hired (on alternating rows, offer data are not available for 2011) against performance on the three selection stages that we examine. Scores are normalized so that zero is equal to the passing cutoff.[14] We

[13] Note that, in 2011, due to the logistical difficulties which delayed auditions, applicants who passed the interview were also placed into the recommended pool.

[14] Appendix Figure 1 shows similar plots of the relation between these scores and whether an applicant passed to the next stage. There are large jumps in the probability of passing at the thresholds, demonstrating a high degree of, albeit not perfect, compliance with our understanding of the TeachDC selection rules. Appendix Figures 2 and 3 show that applicants' characteristics are smooth across these thresholds; we discuss this in more detail below.

see three main takeaways from these figures. First, there are large discontinuous jumps in the probability of job offer and hire at the passing thresholds, with the jump becoming larger as applicants move closer to landing on the recommended list.[15] A natural interpretation is that principals are far more likely to seek out candidates on the recommended list, although it is possible that making (missing) the list could (de-)motivate applicants to seek out DCPS jobs. However, we find this interpretation unlikely given the magnitudes of the jumps at the threshold, which are as large as 40 points.

Second, the relationship between hiring and applicants' scores appears positive on both sides of the cutoff, although it is sometimes nonmonotonic and somewhat weak in comparison with the jumps at the cutoff. For example, as we move from the very bottom of the audition score distribution in 2013 to scores just below the cutoff, offer (hiring) rates go from near zero to 7.1 (8.9) percent but then jump discontinuously to 44.5 (48.6) percent as we move just across the cutoff. This suggests that applicants who score higher in the TeachDC stages may also be more appealing to principals, but these scores are likely to explain only a modest amount of hiring outcomes conditional on whether an applicant made the recommended list.

Third, while DCPS staff working on TeachDC were certainly aware of the cutoffs when evaluating candidates, there is little evidence of excess mass on any particular side of these thresholds. The score distributions become somewhat coarse in later years, but generally we see a smooth density of scores through the cutoffs. Later we present additional tests to support our use of cutoffs to address potential selection bias in our analysis of performance.

[15] In 2011, the jump at interview is larger than audition, consistent with the fact that, as mentioned earlier, applicants passing the interview in 2011 were also placed on the recommended list. In 2012 and 2013 the jump is clearly largest at the passing threshold for the audition.

*2.2 Performance Evaluation (DCPS IMPACT)*

Since 2009, all teachers in DCPS receive a performance evaluation under the IMPACT system. We discuss key features of the system here and refer readers to Appendix D for more detail. A teacher's IMPACT score is a weighted average of several performance measures, which vary depending on the grade(s) and subject(s) to which the teacher is assigned.

A major component of all teachers' evaluations comes from classroom observations. Each teacher is typically observed five times during the year, three times by the school principal and twice by a "master educator," typically a highly experienced teacher, who conducts observations full-time at many schools. Teachers' performance during classroom observations is scored using the district's Teaching and Learning Framework (TLF) rubric.[16] Observers assign scores in several areas of practice that are averaged within observations, and then these composites are averaged across observations.

Other components of the IMPACT rating include student progress on teacher-assessed learning goals (TAS) and the principal's subjective assessments of a teacher's performance outside the classroom, called "commitment to the school community" (CSC) and "core professionalism" (CP). For teachers of math or reading in grades 4 through 10, individual value-added (IVA) measures of effectiveness at raising achievement are also included and account for a large portion of the teacher's overall rating.[17] Teachers have strong incentives to do well on IMPACT; they are dismissed if their score puts them in the lowest of five performance categories and they receive large increases in salary if they excel (see Dee and Wyckoff 2015).

[16] Examples of aspects of practice within the TLF include "explains content clearly", "engages students at all learning levels", "provides students multiple ways to move toward mastery", "checks for student understanding", "maximizes instructional time" and "builds a supportive and learning-focused classroom."

[17] In some years, a school value-added score was included in IMPACT evaluations for teachers without IVA. We do not use this school VA component in our measure of teacher performance, which we describe in Appendix F.

Adnot et al. (2017) present evidence on the validity of IMPACT score as a measure of teacher quality, showing that, at least in grades and subjects with annual tests, the average student test score for a given grade and school increases substantially when a teacher is dismissed from that grade and school because the teacher received low IMPACT ratings.

Finally, it is important to note that DCPS principals have reasonably strong formal incentives to hire effective teachers. Since the school year 2012-13, principal performance in DCPS has been evaluated under the IMPACT system. Like the teacher evaluations, an overall performance score for each principal is generated using multiple criteria such as student test scores and rubric-based evaluations by supervisors. Low scoring principals are dismissed while high scoring principals can receive substantial bonus payments.

### *2.3 Data, Variable Construction and Descriptive Statistics*

We use data on over 7,000 individuals who applied through TeachDC in the recruiting seasons from 2011 to 2013 and who were eligible for a teaching license in DC. We analyze subsequent hiring and performance data from the school years 2011-12 through 2016-17. Thus, we have three cohorts of candidates and new hires, and can observe retention and performance for the 2011 applicants for up to six years. We focus on a small set of key data issues here, while Appendix F provides additional details.

A few limitations in our data are worth noting. First, we have information on job offers but only for 2012 and 2013. As explained below, the offer information helps us to disentangle the effects of labor demand (job offers) and labor supply (offer acceptances), and we find our results are robust to using offer receipt instead of being hired as an outcome.[18] Second, we

[18] For the years 2012 and 2013, 69 percent of TeachDC applicants who received a formal offer end up accepting it; among recommended candidates, 94.8 percent accept. We cannot observe if an applicant receives an informal offer from the school principal before the principal registers the offer formally with the district, and any such informal offers that were turned down were likely not recorded, making these acceptance rates an upper bound.

cannot observe teacher hiring or performance in charter schools. Although they enroll close to half of local students, charter schools are governed by a separate authority, the DC Public Charter School Board. We are also unable to observe if applicants take a job in another school district, such as in Virginia or Maryland. Finally, we do not have student- or classroom-level data from which we might calculate our own measures of teacher performance.

Our primary job performance measure combines all of the IMPACT components using factor analysis. This analysis consistently yields just one significant "performance" factor, which we standardize within school years.[19] We prefer the performance factor over the official IMPACT score because the factor analysis indicates similar weights on each component across years (Appendix Table 2), while there were considerable changes across years in weights used by IMPACT (e.g., the TAS component score, while available, is completely omitted from the calculation of IMPACT for Group 1 teachers in 2011). However, we also present results which examine each component of IMPACT separately and using a standardized version of the official IMPACT score generated by DCPS yields qualitatively similar results (Table 7).

Economists typically focus on value-added as a measure of teacher performance, often because administrative data lack other outcomes. However, there are several reasons why our power is far diminished for examining value-added as a performance measure relative to our use of the combined IMPACT components. First, only about one in five teachers from our main sample has value-added data. Second, value added is typically less stable over time than other

[19] We run factor analyses by year and separately for teachers with and without IVA data (see Appendix Table 2). These are mean zero by construction, and we set each of their standard deviations to one. Underlying the finding of a single factor are the fairly high correlations between different elements of IMPACT. Apfpendix Table 3 shows pairwise correlations for all DCPS teachers from 2011-12 to 2016-17, which are all positive and significant, ranging from 0.63 for the correlation of TLF (class observation) scores and Commitment to School Community to 0.10 for the correlation of Core Professionalism and value-added for math.

performance measures due to a high proportion of noise from test measurement error and idiosyncratic classroom effects (see Hanushek and Rivkin 2010, Ho and Kane 2013). Third, we obtain IVA scores directly from DCPS on a somewhat arbitrary scale from 1.0 to 4.0 that generates additional imprecision.[20] Fourth, DCPS changed its standardized test formats in 2015, and value-added was not included in the IMPACT scores for school years 2014-15 and 2015-16. This means we observe between one and three years of value-added per teacher, as opposed to observing between three and five years of performance more generally. For all of these reasons, our examination of teachers with value-added is far from ideal. We therefore present it after our findings using the broader performance measure for our full sample of teachers.

There are three application scores: pedagogical content knowledge (PCK), interview, and audition. Each is a rescaled composite of the underlying rubric sub-scores assigned by DCPS staff, generated by taking averages or using factor analysis to combine measures when appropriate. All three measures are standardized to have a mean of zero and standard deviation of one within each year of application, and we adjust standard deviations to take account of missing data on applicants who failed to reach the cutoff at a prior stage (see Appendix F).

To simplify many of our regression specifications, we also combine the SAT/ACT, college GPA, Barron's rank, and Master's degree variables into an "academic index" by taking the simple average of each applicant's non-missing variables, standardizing each variable before averaging and then re-standardizing. This index is centered at zero and has a standard deviation of one for the population of applicants, by construction. We calculate a similar "screening index"

[20] We do not have student-teacher linked data to estimate value-added. DCPS (via an outside vendor) calculates IVA using a standard approach that controls for prior student achievement and other covariates (Isenberg and Hock, 2012; Isenberg and Walsh, 2014a, 2014b). Teachers' raw IVA scores are then grouped into a skewed, non-normal distribution with the top and bottom truncated at the extreme values (1.0 and 4.0). Appendix E provides further details. We obtained "raw" IVA scores for 2011-12, but in 2012-13 and 2013-14 we invert the scaled IVA measures using quantiles to generate value-added that is normally distributed, albeit still with truncation in the extremes. The correlation of "raw" and "inverted" IVA for 2011-12 is 0.98, providing reassurance in our transformation.

by summing the (up to three) non-missing standardized screening score values and re-standardizing. Our main results regarding the predictive power of the screening scores for hiring and performance are quite similar if we restrict the sample to applicants with all three scores (Appendix Table 8).

Table 1 presents summary statistics on applicant characteristics, both for the full sample and among those hired into DCPS. Roughly one third of applicants have no prior full-time teaching experience, another third have between one and five years, and the remaining third have more than five years. The proportion of rookie teachers among those hired is somewhat smaller (28 percent). Average self-reported undergraduate GPA (3.4) and composite SAT/ACT scores (1149) are nearly identical regardless of hiring outcome, while college selectivity is slightly higher among hired applicants (2.9 vs. 2.8 on a scale from 1-5).[21] About 12 percent of applicants attended undergraduate or graduate school in Washington DC, but they are overrepresented among those hired (17 percent). Just over half (51 percent) of applicants report having received a degree beyond the BA, and this proportion is slightly greater among the hired (54 percent). Teachers hired in DCPS have an average academic index of 0.07, suggesting potentially modest selection on academic traits. The average screening score index among hired applicants is 0.50, consistent with the patterns in Figure 1.

Table 2 shows pairwise correlations among the background characteristics and selection scores. Academic achievement measures (e.g., undergraduate GPA, SAT/ACT score) all have modest positive correlations, as one might expect. Academic achievement also has small positive correlations with TeachDC scores, particularly the subject-specific written assessment. Prior

[21] The SAT scores reported by our sample are somewhat higher that the national average, which would have been just above 1000 for cohorts who, like most of our sample, graduated high school in the late 1990s and early 2000s. College selectivity is measured with *Barron's Profiles of American Colleges* (2009).

teaching experience has small negative correlations with academic measures (SAT/ACT, Barron's Rank, and undergraduate GPA) and the PCK score, but small positive correlations with having a master's degree and the audition score. Prior experience, academics, and screening scores are essentially uncorrelated with being from the DC (MD-VA) area. Interestingly, while the correlations among the three application scores are positive, they are all small in magnitude, with the highest correlation between the interview and audition scores (0.22). These correlations suggest the potential for each stage in the application process to be capturing distinct information about teaching applicants, rather than repetitively measuring the same set of skills. Low correlations may also indicate a considerable amount of noise in each score.

## 3. Empirical Strategy

Our analysis addresses two related questions: (1) To what extent do applicant characteristics (including screening scores) predict whether an applicant is offered a job and hired? (2) To what extent do applicant characteristics predict teacher performance and attrition?

### *3.1. Offers and Hiring*

We examine both being offered a job and being hired into DCPS as both of these outcomes are important for thinking about our results related to job performance. Job offers are more cleanly interpreted as being driven by labor demand, although it is possible that our formal job offer data do not capture informal offers and we lack offer data from 2011. In addition, the observation of performance is contingent on hiring, rather than job offer, so concerns about

selection bias ultimately hinge on hiring. In practice, our results are quite similar regardless of the outcome used, though there are some differences which we highlight below.

We examine the relationship between applicant characteristics and the likelihood of being offered a job (or being hired) using linear probability models of the form:

$$(1) \qquad H_i = \boldsymbol{X_i}\beta + \boldsymbol{S}_i\gamma + \delta_{jt(i)} + \varepsilon_i$$

where the dependent variable $H_i$ is either an indicator for being offered a job or being hired into DCPS as a teacher, $\boldsymbol{X_i}$ is a vector of background characteristics, and $\boldsymbol{S}_i$ is a vector of scores on the screening assessments. Because not all candidates have complete data for all characteristics, we set missing values to zero and include a set of missing variable indicator flags into the regression. The availability of positions and the supply of candidates may vary by subject area and over time, so we include fixed effects, $\delta_{jt(i)}$, for the "job type," defined as the subject area and grade level for which the applicant applied, interacted with application year. In practice, fixed effects have little impact on our estimates. We also find similar results limiting our sample to applicants with no missing screening scores.

The coefficients of interest are contained in the vectors $\beta$ and $\gamma$, which measure the extent applicant characteristics predict district offers and hiring. However, as seen in Figure 1, applicants who made it to the "recommended pool" were substantially more likely to be hired than others. For this reason, we present additional specifications that include indicators for whether the applicant was placed in this recommended pool.[22] This allows us to examine whether the coefficient on a given predictor is driven by its correlation with placement on the

[22] We interact this indicator with application year. In 2011, the recommended pool was extended to applicants passing the interview stage so, in addition to the indicator for passing the audition stage, we include an indicator for passing the interview for applicants in the 2011 cohort. Thus, the vector includes four mutually exclusive indicators: (i) applicants reaching the recommended pool in 2013, (ii) applicants reaching the recommended pool in 2012, (iii) applicants passing the audition in 2011, and (iv) applicants passing the interview but not the audition in 2011.

TeachDC list of recommended candidates. We base our statistical inferences on heteroskedasticity-robust standard errors. We have also performed these analyses using Logit models, and obtain virtually identical results.

It is worth emphasizing that we do not assign a causal interpretation to our hiring estimates. Characteristics such as undergraduate GPA or performance in an interview are almost certainly correlated with other factors that school officials could observe and value. However, these estimates are nonetheless interesting as a description of hiring practices as well as informative for the interpretation of our analysis of teacher performance.

In addition to selection into DCPS, we are also interested in whether teachers with certain characteristics are systematically hired by certain types of schools within the district. To explore this, we estimate how the characteristics and application scores of newly hired teachers predict the test scores at the schools where the teacher was hired.[23] We estimate a specification similar to Equation 1, except that the outcome variable is the proportion of students scoring proficient or higher on district-wide standardized tests in 2010-11. We also control for the subject and grade level of the teaching position and limit the sample to newly hired teachers.

We also repeat these analyses focusing on applicants offered a job, and achievement at the schools where they were offered that job. We present offers as secondary results because they are quite similar to estimates for hiring and, as mentioned above, we lack offer data for 2011 and must restrict these analyses to the 2012 and 2013 applicants.

*3.2. Performance*

[23] Unfortunately, we do not have measures of school value-added (or principal characteristics) and cannot examine the extent to which more effective schools (or certain types of principals) hire certain types of teachers.

To examine the relationship between applicant characteristics and teacher performance, we must restrict our attention to applicants that were hired by DCPS, for whom we can observe performance. Using this sample, we estimate a series of regressions of the form:

$$(2) \qquad P_{it} = \boldsymbol{X}_i\beta + \boldsymbol{S}_i\gamma + \alpha_{jt(it)} + \theta_{(t-\tau)} + \varepsilon_{it}$$

where $P_{it}$ is the performance of teacher $i$ in school year $t$. The vectors $\boldsymbol{X_i}$ and $\boldsymbol{S}_i$ are as defined above for Equation 1. As in our hiring regressions, missing characteristics are set to zero and missing variable indicators are included in the regression. To account for possible differences in evaluation standards, we include fixed effects for "job type", $\alpha_{jt(it)}$, defined as the subject area and grade level in which the teacher worked during year $t$, interacted with school year. We include additional controls in some specifications to account for possible selection bias (described in Section 3.3 below). We also test robustness to the inclusion of school fixed effects to address potential bias due to non-random sorting of teachers to schools. Last, but not least, since we observe annual performance for each newly hired teacher between one and six times, our sample is an unbalanced panel; accordingly, we include nonparametric controls (fixed effects) for the number of years since hire, $\theta_{(t-\tau)}$ where $\tau$ is the year of hire, and we report heteroskedasticity-robust standard errors that are clustered by teacher.[24]

As in our hiring analysis, we are not seeking to estimate the causal impact of application measures on teacher performance. For example, if individuals with unobservable positive traits (e.g., strong work ethic) sort into selective colleges, then we might find a positive coefficient on college selectivity even if selective college attendance itself has no causal effect on performance. Given the primary purpose of teacher selection, we do not view this as a limitation.

[24] The correlation in performance across adjacent years within teachers in our sample is 0.68, supporting the notion that errors will be correlated within teacher over time. We have estimated models that cluster by school, and by teacher and school, and obtain virtual identical results. Models that restrict the sample to each teacher's first year in DCPS are also qualitatively quite similar, but somewhat less precisely estimated.

*3.3 Accounting for Selection into Hiring*

A primary concern in our analysis is that selection into our sample of hired teachers could bias the estimates in Equation 2. If the correlation between an unobservable performance-related characteristic and applicants' observables (e.g., academic achievement, TeachDC interview score) changes when we condition on being hired, then our estimated coefficients may be biased. Typical stories one could tell suggest negative biases on characteristics positively associated with job offers. For example, principals may make offers based on a combination of characteristics we observe and other teaching skills that are unobservable (to us), so candidates that "look bad on paper" are positively selected on some other dimension. Similarly, if applicants reject offers when they have a combination of better observable and hard-to-observe skills (e.g., because of good outside job opportunities), then applicants who "look good on paper" and still accept will be negatively selected on some other dimension. However, one can easily construct counter examples and it is impossible to sign selection bias *a priori*.[25]

However, as in Goldhaber et al. (2017), we can try to address the issue to some degree using exogenous variation in hiring. Specifically, we can test for selection into DCPS on unobservables by leveraging the sharp discontinuity in the likelihood of hire associated with making it to the TeachDC recommended list, as shown in Figure 1. This discontinuity provides us with an arguably exogenous change in hiring that, conditional on the underlying selection scores, should be independent of unobservable determinants of teacher performance. To support this assumption, Appendix Figure 2 graphs the relationship of applicants' academic index values to the screener scores in each year. This figure reveals no sharp changes in these observable

[25] For example, if applicants reject offers when they have a combination of better observable skills and very little commitment to teaching (where commitment is an unobservable but valuable teaching skill), then those with better observable skills who accept offers will be positively selected.

characteristics coincident with the jump in hiring at the cutoffs, and helps reduce the concern regarding discontinuous changes in unobservable dimensions of teacher quality.[26]

Unlike in a typical regression discontinuity (RD) design, we are not seeking to estimate the causal "impact" of being hired. Instead, we seek to estimate the true correlation between screening scores (which would be the "running variables" in an RD setting) and teacher performance, purged of bias due to non-random selection into teaching. To that end, we use the sharp discontinuity in how the screening scores predict the probability of being hired as an exclusion restriction in a formal, parametric selection correction.

Following the discussion in Heckman and Navarro-Lozano (2004), we first calculate applicants' predicted probabilities of being hired from fitting an augmented version of Equation 1 to which we add instruments based on the screening cutoffs.[27] We then include a polynomial in this predicted probability directly in Equation 2 as a control function meant to account for any potential selection associated with being hired. This is an extension of the traditional approach based on the Inverse Mills Ratio, with the identifying assumption that the instruments are associated with the likelihood of being hired, and thus having an outcome measure, but do not influence the outcome directly. For robustness and transparency, we also show results from an additional approach where we simply include indicators for being in the recommended pool directly in Equation 2. As we show below, these indicators are not significantly associated with teacher performance (conditional on the other variables in the model, such as the continuous

[26] Out of the nine panels in Appendix Figure 2, the one with the greatest apparent change in average academic index at the cutoff is based on audition scores in 2011. But there is no change in hiring probability at this cutoff because applicants who passed the interview in 2011 were also put on the recommended list (see Figure 1). Appendix Figure 3 shows continuity across cutoffs for applicants' prior experience, age, and probability of being from the DC area.

[27] The instruments are two indicator variables: (i) Applicants in any year who scored above the audition cut-score designated by DCPS as the threshold for the recommended pool. (ii) Applicants in 2011 who scored above the interview cut-score, who were also placed in the recommended pool as discussed in the text.

screening measures), so it is not surprising that our results are highly robust to these alternatives for addressing selection bias.

Note that there are two important limitations of this approach. First, it informs us about the presence of bias in the relationship between screening scores and performance arising from selective offers (or acceptances of offers) made to candidates with screening scores near the cutoff for the recommended pool. If offers are only being made/accepted selectively (based on unobservables) for candidates with scores far from the cutoff, then this RD approach is less informative.[28] Second, this strategy does not directly inform us about selection biases in other characteristics (e.g., prior experience) due to non-random hiring. However, a lack of evidence for bias in the selection scores due to non-random hiring provides some assurance that any biases in other coefficients are likely to be quite small.

## 4. Results

### *4.1 Teacher Hiring*

Estimates of Equation 1 are presented in Table 3, beginning with regressions where each characteristic is entered separately (Columns 1 and 2), and then simultaneously (Columns 3-5). Several robust patterns emerge. For the most part, applicants' academic credentials appear to bear little relation to being hired into DCPS. In Column 1, the coefficients on undergraduate GPA, SAT/ACT score, and master's degree are all close to zero and statistically insignificant, while the coefficient on college selectivity is positive but exceedingly small (less than 1 percentage point for each point on the 1-5 Barron's scale). Moreover, when we add controls for reaching the recommended pool (Column 2), the coefficients all shift downward, reflecting the

[28] Goldhaber et al. (2017) face a similar issue; their IV is the presence of an error in the calculation of candidates 1st round score that causes a candidate to cross the cutoff needed to be scored in the 2nd round.

small positive correlation between academic background and screener performance seen in Table 2. The coefficients on college selectivity and master's degree now become slightly negative and insignificant, and those for undergraduate GPA and SAT/ACT score become negative and statistically significant, though small in magnitude (about 1-1.5 percentage points). Pooling these measures into an academic index yields the same pattern: a very small positive coefficient (0.005) without conditioning on making it to the recommended pool or not, and a very small negative coefficient (-0.012) when these controls are added.

These results are consistent with two interpretations. First, school principals may not put positive weight on these basic academic achievement measures when making hiring decisions, consistent with some prior work (Ballou 1996, Hinrichs 2014). Second, it may be that principals do place positive weight on these characteristics, but that applicants with better academic backgrounds are less likely to accept job offers from DCPS schools. We cannot definitively separate supply and demand explanations given the limitations in our data. We do, however, observe many offers that were declined – one-third of offers in 2012 and 2013 did not result in a hire – permitting a partial empirical test. In Table 4 we present results parallel to Table 3 separately for the outcome "hired" and the outcome "offered a job," restricted to the 2012 and 2013 applicants from whom we have offer data. The pattern of point estimates is similar regardless of whether offer or hire is the dependent variable, supporting the demand-side interpretation.[29]

Not surprisingly, we find that each of the three application scores is positively associated with the likelihood of being hired when we do not control for reaching the recommended pool (Table 3 Column 1), with coefficients rising monotonically as we move to the later stages of the

[29] In Appendix Table 4 we provide a complete parallel to Table 3 for the outcome of receiving a job offer.

TeachDC selection process. A one standard deviation increase in applicant score is associated with increases in the likelihood of being hired of 6.0, 10.8, and 15.8 percentage points for the PCK, interview, and audition, respectively.

These effects are quite large, given the baseline hiring rate of roughly 13 percent, but is likely be driven by the effect of arriving into the recommended candidate pool. When we include fixed effects for reaching the recommended pool (Table 3, Column 2), the coefficient on the PCK written test goes essentially to zero, while those on the interview and audition drop by at least two-thirds. Pooling the three scores into a selection index yields a similar pattern, with a coefficient of 0.069 dropping to 0.023 when we control for being in the recommended pool. Again, in Table 4 we show that the coefficients on the three application scores are quite similar when the outcome is offer instead of hire, consistent with a demand-side interpretation of the results. These results suggest that principals did not rely heavily on the screener scores collected by TeachDC, despite relying on whether applicants were recommended, and that whatever factors principals did rely on were not highly correlated with these scores.[30]

We also find that applicants with no prior teaching experience are less likely to be hired by DCPS schools than individuals with prior experience. Depending on the comparison group and specification, rookie applicants have roughly a 3-5 percentage point lower probability of being hired in DCPS.[31] In considering the role of experience in hiring, it is important to note that

[30] One might hypothesize that principals did examine scores of applicants on the list, which they could potentially access online, but did not rely on (anything correlated with) scores for other applicants. We explore this by running specifications that interact the screener scores with an indicator for being on the recommended list (available upon request). We find no evidence to support this idea; the coefficients are quite similar for interview and audition, and the PCK coefficient is somewhat larger for applicants who did not make it to the recommended pool.

[31] Applicants that report more than ten years of prior experience are only slightly more likely to be hired than those reporting just one or two years, but when we limit the sample to 2012 and 2013 when we observe job offers (Table 4), the relationship between experience and hiring is stronger and significant for all categories. There is also some evidence that applicants with more than two years of prior experience are more likely to turn down offers than those with two or fewer years.

the financial burden of paying higher salaries to teachers with more experience is borne by the school district, not individual schools.

Finally, local applicants from Washington DC and, to a lesser extent, Maryland and Virginia, are more likely to be hired.[32] These coefficients are stable across specifications and suggest an increase of about 5-6 percent for DC applicants and 2-3 percent for applicants from the greater DC area, relative to applicants from farther away. This is consistent with Hinrichs' (2014) national teacher resume audit study, where interview requests occurred about 35 percent less frequently for "applicants" whose CV indicated having attended an out-of-state college.

Teaching positions may vary in requirements (e.g., math skills for high school physics versus elementary art), and it is natural to ask if the coefficients in Table 3 might vary considerably across subjects and grade levels. Appendix Table 5 shows results akin to Column 5 of Table 3, where we split the sample into four groups based on applicants' self-reported area of interest: elementary and early childhood education, middle and high school core subjects, special education, and other specialties.[33] Broadly speaking, we find little evidence of heterogeneity in the relationship of applicant characteristics to hiring. Academic index coefficients are small and negative, screener score index coefficients are small and positive, and being a local applicant boosts hiring probability. Experience coefficients are all near zero for applicants in Other Specialties (e.g., arts, foreign languages, ELL teachers) but not very precisely estimated, and they are somewhat larger for applicants in Elementary and Early Childhood Education.

[32] Recall that we do not have applicants' addresses and use their college/university location as a proxy.
[33] "Middle/High school core subjects" includes English, math, sciences, and social studies. "Special education" includes teachers specializing in special education at any grade level. "Other specialties" includes arts, foreign languages, physical education and health, and English language learners at any grade level.

*4.2 Sorting of New Hires*

Table 5 presents estimates of sorting across schools with different levels of student achievement for newly-hired teachers from TeachDC, and similarly for TeachDC applicants offered a job. Teachers with better academic backgrounds are more likely to be both offered a job and hired by schools with higher student achievement levels. A new hire with a one standard deviation higher academics score is, on average, working in a school with 4 percentage points more students scoring proficient or advanced on district tests; about one-fifth of a standard deviation in the school achievement distribution. The same pattern is true when we examine job offers made to teachers, suggesting this sorting is due in large part to school demand, not just applicant preferences. We find little evidence of sorting on screening scores and some suggestive evidence that teachers with more than 5 years of experience are slightly less likely to receive job offers from higher achieving schools.[34]

Together these results suggest that there is sorting of teachers within DCPS based on applicant characteristics, and that it could contribute to inequality in student outcomes within the district. Applicants' academic credentials are associated with teaching in higher achieving schools. This type of "regressive" sorting is consistent with findings from studies of transfer behavior among experienced teachers (Lankford et al. 2002, Hanushek et al. 2004, Jackson 2009). As mentioned above, when we examine teacher performance below, we explore robustness to the inclusion of school fixed effects to address the issue of non-random sorting.[35]

[34] The estimates in Table 5 use hires and offers from 2012 and 2013, the years when we have offer data. We find very similar results if we estimate Columns 1 and 2 using the entire sample from 2011 through 2013. The cross-sectional correlation of teacher experience (as proxied by years since hire) and school achievement is small and positive (about 0.1), so the pattern we find for TeachDC applicants' prior experience does not seem to reflect broader sorting patterns (e.g. retention and transfer) within the district.

[35] Table 5 suggests principals at higher achieving schools are more likely to hire (offer jobs to) teachers with stronger academic backgrounds. This may be rational behavior if teachers' academics have a higher return in higher achieving schools. In Appendix Table 7 we test this hypothesis building on the performance regressions presented

*4.3 Teacher Performance*

We now restrict our attention to applicants for whom we can observe performance, and we estimate regressions based on Equation 2, with job performance as the dependent variable. The main results are presented in Table 6. Overall, the patterns we find are strikingly different than those discussed earlier for the outcome of being hired (Table 3) or receiving a job offer (Table 4).

The academic index did little to predict hiring outcomes—indeed its estimated effect was slightly negative once we control for being in the recommended pool—but it is a strong positive predictor of performance, with an effect size of over 0.2 standard deviations.[36] The effect size for the screener score index is similar in magnitude, while the coefficients for applicants' prior experience and being from the DC region have inconsistent signs and are almost all statistically insignificant.[37]

When we test the sensitivity of these coefficients to including controls for being in the recommended pool, we also see an important departure from the hiring results. While the academic index flipped signs (from small and positive to small and negative) in the hiring regressions, the coefficient is extremely stable (0.210 to 0.211) with or without the recommended pool controls. The coefficient on the screener index is also quite stable (0.212 to 0.220) when we

later in Table 6. In regressions predicting teacher job performance we interact academic index with school achievement level (the outcome variable in Table 5). The interaction coefficient is negative but far from statistically significant, and turns slightly positive if we limit estimation to within school variation.

[36] We focus on the indices of academics and screener scores, but results with disaggregated covariates (Appendix Table 6) are all qualitatively similar.

[37] Readers familiar with the value-added literature will note that studies typically find a positive association between experience and value-added, particularly in the early years of a teacher's career. While the results on prior experience in Table 6 seem to contrast with this finding, studies linking value-added and teacher experience are usually based on variation within teachers (and within a district) over time. As mentioned earlier, we do find large improvements in performance over time within DCPS for the same teacher. Also, when we focus on teachers with value-added scores (Table 9), there is a stronger positive relationship between experience and our broader teacher performance measure. This suggests that the experience-performance relationship documented in the value-added literature may be stronger for teachers in commonly tested grades and subjects.

control for the recommended pool, contrasting with hiring where the coefficient shrank by roughly two-thirds (0.069 to 0.023).

The reason we see so much stability is because the controls for being in the recommended pool simply do not predict performance (the F-statistic is 0.83), despite being highly significant predictors of hiring (F-statistic of over 75).[38] Thus, the variation in academics and screening evaluations among candidates *within* recommended and non-recommended pools strongly predicts performance, even though principals appear to place much more weight on "between-pool" variation when hiring new teachers. Similarly, estimates are also quite stable when we implement the control function approach from Heckman and Navarro-Lozano (2004) to correct for selection bias. The excluded instruments are highly significant predictors of being hired (F-statistic 42.41), but controls for the predicted probability of hire are not significant predictors of performance. The academic index coefficient is somewhat smaller (0.15 rather than 0.21) as is the screener index coefficient (0.17 rather than 0.21). Coefficients are negative but small and insignificant for differences in performance between those who are from the DC area and elsewhere. These estimates are all insensitive to our selection corrections. Overall, this provides strong support for the notion that selection is not driving our results.

We showed earlier (Table 5) that the relationship between applicant characteristics and hiring varies somewhat across schools, and one might be concerned that evaluations in some schools are spuriously higher than others. By adding school fixed effects to the performance regressions (Table 6 Column 4), we base identification on comparisons of applicants hired into the same DCPS school. These controls address some concerns about omitted variables (e.g.,

[38] For reassurance on this issue, we plot average performance against the screener scores for each score and year in Appendix Figure 4. The performance - screener relations are all smooth.

some principals are "easier graders"), but research suggests teachers do sort across schools based on "true" quality, and within-school comparisons may exacerbate other sources of bias.[39] Regardless, school FE do not change the basic results. While they capture a significant amount of performance variation (F-statistic of 13), the coefficients on academics (0.18) and screening scores (0.17) are only slightly attenuated and remain statistically and economically significant. We find very similar results if we use school-year FE, which would also account for principals changing over time.

Finally, it is worth noting again that we use a single measure of job performance because our factor analysis suggests that the components of IMPACT scores primarily reflect a single underlying metric. But the performance incentives for teachers in DCPS relate to the actual IMPACT score, and it is possible that the results are driven more by some pieces of IMPACT than others. We therefore presents results using the specification from Table 6 Column 2, where we examine the official IMPACT score and each IMPACT performance component in a separate regression (Table 7).[40] The coefficients on the academic and screener indices are positive and highly significant for all score components, suggesting that they capture a wide variety of skills being evaluated and that our results are robust to component weighting.[41] We do see some evidence that teachers with between one and five years of experience perform better on classroom evaluations (both independent and internal) than those with no prior experience; they do not, however perform differently in terms of core professionalism, broader contributions to the school, and teacher-assessed student learning. We see no evidence of performance criteria positively related to being from the DC area.

[39] See Jackson (2009) for evidence on teacher quality and sorting. Attenuation bias (i.e., teachers with bad observables must have good unobservables since they were still hired) is likely even more salient within a school.
[40] As discussed earlier, value added is available for a small subset of our sample, which we address in Section 4.4.
[41] For interested readers, we also present results using the actual IMPACT score (standardized) in Table 7.

To summarize these results and get a better sense of magnitudes, we estimate predicted first-year performance using applicants' characteristics and plot kernel densities of *actual* performance by *predicted* quartile in Figure 2.[42] Applicants from the top *predicted* quartile perform far better *in practice* than those from the bottom quartile. Indeed, the teachers in the top quartile of predicted performance score more than two-thirds of a standard deviation (0.71) higher on actual performance compared with teachers from the bottom quartile of predicted performance. To provide some context, the average improvement in performance between a teacher's first and third year –when we might expect significant performance gains – is just 0.37 standard deviations among all DCPS teachers during our sample period.[43]

We can also use the predicted performance of applicants to summarize the influence of the current application process on hiring decisions. Figure 3 plots distributions of predicted first-year performance separately for four groups of applicants, defined by whether they were on the recommended list and whether they were hired.[44] Note that predicted performance is a (regression) weighted average of applicant characteristics; we cannot observe what the actual performance would have been for applicants that were not hired. The distributions of predicted performance are much higher for recommended candidates, not surprising given the results in Table 6. Moreover, there is virtually no difference between the predicted performance

[42] The estimates are based on coefficients estimated with the specification in Appendix Table 6 Column 2, limiting the sample to TeachDC applicants' first year teaching at DCPS. Appendix Table 6 Column 2 is the same as Table 6 Column 1 except that the subcomponents of the academic index and screening scores index are included as separate regressors. The subject-taught by year fixed effects are not included in our predictions. We use a leave-one-out procedure – i.e., to obtain the predicted value for each teacher *i*, we estimate our model using all observations except for those from teacher *i* — so a teacher's outcome does not influence his or her own predicted score. Recall that the performance variable is standardized based on the population of DCPS teachers within school years.

[43] Simulations of teacher hiring in Staiger and Rockoff (2010) and Rothstein (2015) assume performance gains of 0.47 and 0.33 teacher standard deviations, respectively, for teachers between years 1 and 3. So the finding for our sample is fairly similar to the broad conclusions of observers of this literature.

[44] Appendix Figure 5 plots applicants' predicted performance by whether they were on the recommended list and whether they were ***offered*** a job, using only 2012 and 2013. The results are the same: higher predicted performance for applicants on the recommended list but no noticeable differences between those offered and not offered a job, regardless of whether the candidates were on the list.

distributions of those hired or not hired from the recommended pool, again suggesting principals did not heavily weight academic background or screening scores within the recommended pool. There are also no differences by hire status for applicants who did not reach the recommended pool, suggesting (correlates of) these predictive characteristics are also not weighted heavily by principals when hiring through other channels.

Last, but not least, we offer an admittedly speculative "back-of-the-envelope" calculation of the gain in teacher performance if DCPS had hired TeachDC applicants in descending order of predicted performance rather than the way applicants were actually hired. This calculation assumes (1) the number of TeachDC hires in each year did not change and (2) the correlation of observables and unobservables that predict performance is the same among hired and non-hired applicants. We find that average performance would have risen by 0.42 standard deviations. 74.9% of applicants who were actually hired would not have been hired under this hypothetical "descending order" system.

*4.4 Heterogeneity Across Sub-Populations and Teachers with Value-Added*

As noted in the introduction, one of the advantages of our setting is that we can estimate performance regressions for teachers in all grades and subject areas. We therefore have an opportunity to explore heterogeneity in our performance predictions along these dimensions. We first divide the sample into four groups defined by grade level and subject area in which the new hire is working: elementary and early childhood education, middle and high school core subjects, special education in any grade, and other specialties.[45] As reported in Table 8, the relationships

[45] These are the same four groups we used to examine heterogeneity in hiring, but are defined here by subject and grades taught instead of subject and grades to which candidates applied. "Middle/High school core subjects" includes English, math, sciences, and social studies. "Special education" includes teachers specializing in special education at any grade level. "Other specialties" includes arts, foreign languages, physical education and health, and English language learners at any grade level.

between performance and academic background or screening scores are quite similar across these four groups and always statistically significant. Estimates for experience and being from the DC area exhibit no clear patterns across the groups but are somewhat imprecise.

Finally, we examine the subsample of roughly 200 math and English teachers in grades four through ten who have individual value-added (IVA) scores.[46] The results are shown in Table 9. While the IVA results lack precision for a variety of reasons (see Section 2.3) there are some interesting patterns. The academic and screener indices are predictive of overall performance for the set of teachers with value-added (Column 1), with coefficients of just over 0.2 standard deviations, similar to the full sample analyzed in Table 6. When we add school fixed effects, the results when examining overall performance are smaller and less precisely estimated (Column 2), following the pattern in the full sample (Table 6) but far less robust. When we turn to IVA, the coefficient on academic index is significant and positive but this is greatly attenuated when we add school fixed effects (from 0.19 to 0.03). The coefficient on the screener index is negative (-0.11) using both between and within school variation but positive (0.10) with school fixed effects, and neither point estimate is statistically significant.

Table 9 also shows some evidence that performance is increasing with prior teaching experience for this subset of teachers, both for the overall metric and for value-added, although the point estimates are not always monotonic nor significant at conventional levels. However, this pattern is more consistent with prior studies showing positive returns to experience for teachers in these grades and subjects (e.g., Rockoff 2004, Papay and Kraft 2015).

Given the primacy of value-added as a measure of teacher performance in the economics literature, these results are an important caution to our primary findings. Yet the standard errors

[46] Math value-added scores are only available in grades 4-8. ELA value-added scores are available in grades 4-10 in 2012-13, 2013-14, and 2016-17, but only grades 4-8 in 2011-12.

are roughly 2-3 times larger when we examine value added for this subsample, compared to examining overall performance for the full sample of new hires, and we cannot rule out effects of the size documented by Goldhaber et al. (2017).[47] We therefore view our setting as providing a weak test of whether applicant characteristics can predict performance on IVA, and more data will need to be brought to bear on this question. We remain confident in our main conclusions, particularly given recent studies documenting that sizeable effects of teachers on other "non-cognitive" outcomes are only weakly correlated with teacher value-added (Jackson 2016, Petek and Pope 2017).

*4.5 Attrition*

Hiring effective teachers will be more beneficial when they stay employed in the school or district for a significant time period. In our setting, given the availability of jobs at charter schools and the suburbs surrounding Washington DC, one might be concerned that applicants with good academic credentials or those that score well on the TeachDC screeners could end up leaving the district quickly. We therefore examine attrition from DCPS and from the first school in which new hires were employed (see Table 10). We can observe four years of attrition for the 2013 cohort and show how attrition evolves over a four-year window pooling the three cohorts. The mean four-year attrition rates for teachers in our sample are 57 percent for leaving DCPS entirely and 73 percent for leaving the school that hired them initially.

Academic background shows a small positive relationship to attrition after the first year. Teachers with one standard deviation better academics scores are about 4 percent more likely to

[47] To compare our results with those from Goldhaber et al. (2017) we need to convert to standard deviations of student-level test scores. One standard deviation in the IVA distribution is roughly 0.2 student-level standard deviations, so, for example, the standard error of 0.085 for the screening index (Column 3) translates to a standard error of 0.085*0.2 = 0.017 standard deviations in the student test score metric. This is roughly the same magnitude as point estimates reported by Goldhaber et al. for district screening (0.024 in reading, 0.032 in math) and roughly 30% of the largest estimate they report (0.062 for school-level screening in math).

leave DCPS after the first year, compared to a baseline rate of 20 percent. The coefficient on academic score does not change much in the next three years (we cannot reject equal coefficients across columns 1-4), suggesting that, after the first year, teachers with better academic backgrounds are no more likely to leave the district. Teachers with better academic records are also more likely to leave their hiring school after year one, but four years after hire differences predicted by academic scores are gone suggesting attrition by those with poor academic backgrounds caught up. The coefficients on screener scores are negative – higher scoring applicants are less likely to attrit – and somewhat less precisely estimated, but they follow the same patterns. Higher screening scores predict a lower likelihood of leaving DCPS after the first year (about 2% less attrition), but the effect is not significant and there are no signs of further differential attrition over time. Screening scores predict significantly less attrition from the hiring school after year one (about 4% lower probability of leaving for a one standard deviation increase in the score), but the coefficient shrinks to nearly zero by year four.[48] Experience is a weak predictor of attrition, though teachers with more than five years of prior experience appear slightly more likely to leave their initial school.

The most robust finding from this analysis is that having attended undergraduate or graduate school in Washington DC is strongly negatively related to attrition. These "local" teachers were more than 20 *percentage points* less likely to leave DCPS in the first four years and roughly 16 percentage points less likely to leave their initial school during the same period. Those who attended schools in Maryland or Virginia are about 8 percentage points less likely to

[48] We estimate an additional specification to address the fact that teachers who perform poorly under the IMPACT system are forced to leave DCPS. This could generate mechanical effects on attrition, but these departures might not be deemed costly from the point of view of school principals and DCPS officials. We include an indicator variable for teachers ever given an IMPACT rating of "Ineffective" or rated "Minimally Effective" for two years in a row, both of which trigger dismissal. The results of this specification are quite similar, with the coefficient on selection scores somewhat attenuated toward zero, suggesting part of the unconditional estimate was driven by these teachers being less likely to get a poor IMPACT evaluation.

leave DCPS during this period but no less likely to leave their school than teachers who did not attend school in the DC-MD-VA area. Earlier we found that DC area applicants were more likely to be hired. Since principals oversee hiring and pay the direct (time) costs associated with this task, they may place considerable weight on the probability a new hire remains at the school.

The attrition results suggest that a principal wishing to maximize performance over several years would give higher priority to attractive applicants (e.g. those with strong academic backgrounds and high screener scores) from the DC area. Moreover, if teacher performance improves with experience, hiring teachers who stay longer may, all else equal, improve teacher quality. Since our performance regressions control for the number of years a teacher is in DCPS, they may not be giving local teachers "credit" for staying longer. However, if we omit these DCPS experience controls, we find the coefficients for applicants from the DC area increase by 0.01 standard deviations but remain statistically insignificant, suggesting this connection from attrition to effectiveness is relatively unimportant.[49]

## 5. Discussion and Conclusions

We study applicant characteristics, hiring outcomes, and on-the-job performance of teachers in Washington DC Public Schools (DCPS). In contrast to prior work, our analysis includes teachers in all grades and subjects and a work quality metric based on a variety of performance indicators. We also address issues of selection into employment – and observability of performance – using idiosyncrasies in the placement of applicants into a pool of recommended candidates that are much more likely to get hired.

[49] This is in line with Kane et al. (2008), who find that "even high turnover groups (such as Teach for America participants) would have to be only slightly more effective in each year to offset the negative effects of their high exit rate." Being from the DC area is also uncorrelated with other attributes (Table 2), so conditioning on prior experience, academics, and screening scores in our performance regressions is not masking a "reduced form" effect of being from DC.

We find that academic background and scores on job screening tests strongly predict teacher performance. These results are highly robust to correcting for selection into employment and for non-random sorting of teachers to schools. Using our results, we calculate that the gap in job performance between top quartile and bottom applicants is almost two-thirds of a standard deviation, or 25 percentile points in the distribution of all DCPS teachers.

We also find that these "high quality" candidates are more likely to be hired because of their higher probability of being placed on a recommended list, but district principals do not otherwise appear to seek out applicants with these characteristics. This suggests that there exists considerable scope for improving district teacher quality through the selection process. During the three years we examined, DCPS hired 982 teachers who applied through TeachDC. Yet among the nearly 6,500 TeachDC applicants *not* hired, 764 would be predicted to have first-year performance in the top quartile of the hired teacher distribution. In other words, there were more than enough top quartile applicants not hired to replace the bottom three quartiles (737 teachers) of those who were hired. If these predictions are accurate, this replacement would raise average first-year teacher performance by 0.42 standard deviations.

This paper examines whether highly detailed assessments of applicants' backgrounds and skills have the potential to improve hiring decisions for teachers.[50] This new empirical

[50] We focus on the benefits of improving selection, but evaluating applicants does create additional costs. However, these screening costs appear to be small relative to the potential gains from improved selection, via benefits to students of more effective teachers (Chetty et al. 2014a,b, Jackson 2014, 2016) or the costs of dismissing ineffective teachers after hiring (Rothstein 2015). We approximate total marginal costs for TeachDC to be between $70-200K per year, or between $370-1,070 per new hire. The primary marginal cost is the labor of the DCPS staff who conduct and score the interviews and teaching auditions. Each administrator spends about one hour per interview and one hour per audition. DCPS budgets $63K per year for interviews and auditions at $34 per hour of administrator work. The screening steps leading up to the interviews and auditions are much less costly at $7,500 per year total. In addition, DCPS budgets $133K per year for the staff who manage recruitment and screening. Some of that management cost would be required even without the additional screening measures, so the marginal cost is something less than $133K. This cost per new hire divides the total cost by 190 new hires—the average annual number of new hires during the study period who completed all stages of the TeachDC screening process. The per hire cost would be lower if we count new hires who completed only part of the TeachDC process, but, as discussed earlier, it is difficult to know exactly how the TeachDC measures influenced those hires.

evidence is an important contribution to the small but growing economics literature on employee selection, and our use of a high-stakes performance measure (similar to ones being used throughout the country) provides a novel contribution to the substantial literature on teacher productivity.

An important caveat is that we cannot measure student learning directly for the majority of teachers, and our estimates for teachers with value-added scores are weak and imprecise. Our observation-based performance measure is likely driven, at least in part, by factors that affect student learning, but the correlation between these scores and applicant characteristics could theoretically be driven by factors that are orthogonal to learning (e.g., people who smile a lot get higher audition/observation scores but are not better teachers). More research is needed to test, in a direct manner, the connection of teacher applicant characteristics to student learning in a wide variety of subjects and grade levels.

"Hire the right employees" is an intuitive goal for managers in all sectors, including school principals. How to go about improving hiring is much less clear. We believe that districts would do well to collect applicant data in a more systematic manner, as DCPS has done, to understand which candidates are hired and how they perform. In addition, they should explore ways to provide useful applicant information to school principals, who ultimately make personnel decisions but may have limited bandwidth to gather and analyze applicant and teacher performance data. We plan to study the impact of such information in future work.

## References


Adnot, Melinda, Thomas Dee, Veronica Katz, and James Wyckoff. 2017. "Teacher turnover, teacher quality, and student achievement in DCPS." *Educational Evaluation and Policy Analysis* 39, no. 1: 54-76.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." *Quarterly Journal of Economics* 131(3): 1415-1453.

Ashraf, Nava, Oriana Bandiera, and Scott S. Lee. 2016. "Do-gooders and Go-getters: Selection and Performance in Public Service Delivery." Unpublished Manuscript.

Bacher-Hicks, Andrew, Mark J. Chin, Thomas J. Kane, and Douglas O. Staiger. (2017). "An Evaluation of Bias in Three Measures of Teacher Quality: Value-Added, Classroom Observations, and Student Surveys." National Bureau of Economic Research, Working Paper, No. 23478,

Baker, Al. and Marc Santora. 2013, January 18. "No Deal on Teacher Evaluations; City Risks Losing $450 Million." *New York Times*, p. A1.

Barnes, Gary, Edward Crowe, and Benjamin Schaefer. 2007. *The Cost of Teacher Turnover in Five States: A Pilot Study*. Washington, DC: National Commission on Teaching and American's Future.

Ballou, Dale. 1996. "Do Public Schools Hire the Best Applicants?" *Quarterly Journal of Economics* 111(1): 97-133.

Balter, Dana and William D. Duncombe. 2005. "Teacher Hiring Practices in New York State Districts," Report prepared for the Education Finance Research Consortium.

*Barron's Profiles of American Colleges*, 28th Edition. 2009. Hauppauge, NY: Barron's Educational Series.

Blazar, David and Matthew A. Kraft. 2017. "Teacher and Teaching Effects on Students' Attitudes and Behaviors." *Educational Evaluation and Policy Analysis.* 39(1): 146-170

Boyd, Donald, Lankford, Hamilton, Loeb, Susanna, Rockoff, Jonah and Wyckoff, James (2008). "The narrowing gap in New York City teacher qualifications and its implications for student achievement in high- poverty schools." *Journal of Policy Analysis and Management*. 27(4): 793-818.

Boyd, Don, Hamp Lankford, Susanna Loeb, Matthew Ronfeldt, and Jim Wyckoff. 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.

Brown, Meta, Elizabeth Setren, and Giorgio Topa. 2016. "Do Informal Referrals Lead to Better Matches? Evidence from a Firm's Employee Referral System." *Journal of Labor Economics* 34(1): 161-209.

Burks, Stephen V., Bo Cowgill, Mitchell Hoffman, and Michael Housman. 2015. "The Value of Hiring Through Employee Referrals." *Quarterly Journal of Economics* 130(2): 805-839.

Cantrell, Steven, Jon Fullerton, Thomas J. Kane, and Douglas O. Staiger. 2008. “National board certification and teacher effectiveness: Evidence from a random assignment experiment,” NBER Working Paper 14608.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593-2632.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff, 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review* 104(9): 2633-2679.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor, 2007. “Teacher credentials and student achievement: Longitudinal analysis with student fixed effects,” *Economics of Education Review*, 26(6): 673-682.

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor. 2010. “Teacher credentials and student achievement in high school a cross-subject analysis with student fixed effects.” *Journal of Human Resources* 45(3): 655-681.

Dee, Thomas, and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34(2): 267-297.

Garret, Rachel and Matthew P. Steinberg (2015). “Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence from the Randomization of Teachers to Students.” *Educational Evaluation and Policy Analysis*. 37(2): 224–242.

Goldhaber, Dan., Cyrus Grout, and Nick Huntington-Klein. 2017. “Screen Twice, Cut Once: Assessing the Predictive Validity of Teacher Selection Tools.” *Education Finance and Policy* 12(2): 197-223

Greenberg, Julie, Arthur McKee, and Kate Walsh. 2013. *Teacher Prep Review: A Review of the Nation’s Teacher Preparation Programs*. Washington, D.C.: National Council on Teacher Quality.

Grossman, Pam, Julie Cohen, Matthew Ronfeldt, and Lindsay Brown. 2014. “The Test Matters: The Relationship between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment.” *Educational Researcher 43* (6): 293–303.

Hanushek Eric A., John F. Kain, and Steven G. Rivkin. 2004. “Why Public Schools Lose Teachers.” *Journal of Human Resources* 39(2): 326-354.

Hanushek, Eric A., and Steven G. Rivkin. 2010. “Generalizations about using value-added measures of teacher quality.” *American Economic Review Papers and Proceedings* 100(2): 267-271.

Harris, Douglas N., William K. Ingle, and Stacey A. Rutledge. 2014. “How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures.” *American Educational Research Journal* 51(1): 73-112.

Heckman, James and Salvador Navarro-Lozano. 2004. “Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models.” *Review of Economics and Statistics* 86(1): 30-57.

Hinrichs, Peter. 2014. "What Kind of Teachers Are Schools Looking For? Evidence from a Randomized Field Experiment." Federal Reserve Bank of Cleveland Working Paper 14-36.

Ho, Andrew D., and Thomas J. Kane. 2013. "The Reliability of Classroom Observations by School Personnel," Research Paper. MET Project. Bill & Melinda Gates Foundation.

Hoffman, Mitchell, Lisa B. Kahn, and Danielle Li. 2015. "Discretion in Hiring." NBER Working Paper 21709.

Isenberg, Eric, and Heinrich Hock. 2012. *Measuring School and Teacher Value Added in DC, 2011-2012 School Year: Final Report.* Mathematica Reference Number 06860.501. Washington DC: Mathematic Policy Research.

Isenberg, Eric, and Elias Walsh. 2014a. *Measuring Teacher Value Added in DC, 2012-2013 School Year: Final Report.* Mathematica Reference Number 06838.502. Washington DC: Mathematic Policy Research.

Isenberg, Eric, and Elias Walsh. 2014b. *Measuring Teacher Value Added in DC, 2013-2014 School Year: Final Report.* Mathematica Reference Number 40379.503. Washington DC: Mathematic Policy Research.

Jackson, C. Kirabo. (2016) "What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes," NBER Working Paper No. 22226.

Jackson, C. Kirabo. 2014. "Teacher Quality at the High-School Level: The Importance of Accounting for Tracks." *Journal of Labor Economics* 32(4): 645-684.

Jackson, C. Kirabo. 2009. "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation." *Journal of Labor Economics* 27(2): 213-256.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Research Paper. MET Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City." *Economics of Education Review* 27: 615-631

Kane, Thomas J., and Douglas O. Staiger. 2005. "Using Imperfect Information to Identify Effective Teachers." Unpublished manuscript, April 2005.

Kane, Thomas J., and Douglas O. Staiger. 2011. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Research Paper. MET Project. Seattle, WA: Bill & Melinda Gates Foundation.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying effective classroom practices using student achievement data" *Journal of Human Resources 46* (3): 587-613.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff. 2015. "Value-added modeling: A review." *Economics of Education Review* 47: 180-195.

Lankford, Hamilton, Susanna Loeb, and James Wyckoff. 2002. "Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis." *Educational Evaluation and Policy Analysis* 24(1): 37-62.

Liu, Edward, and Susan M. Kardos. 2002. *Hiring and Professional Culture in New Jersey Schools.* Cambridge, MA: Project on the Next Generation of Teachers at the Harvard Graduate School of Education.

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. "The Validity of Employment Interviews: A Comprehensive Review and Meta-Analysis." *Journal of Applied Psychology* 79(4): 599-616.

Milanowski, Anthony. 2004. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education 79* (4):33–53.

Ost, Ben. 2014. "How Do Teachers Improve? The Relative Importance of Specific and General Human Capital." *American Economic Journal: Applied Economics* 6(2): 127-51.

Oyer, Paul, and Scott Schaefer. 2011. "Personnel Economics: Hiring and Incentives." In *Handbook of Labor Economics 4,* David Card and Orley Ashenfelter editors. 1769-1823.

Papay, John P., Matthew A. Kraft. 2015. "Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement." *Journal of Public Economics* 130: 105–119

Petek, Nathan, and Nolan G. Pope. 2017. "The Multidimensional Impact of Teachers on Students," University of Chicago Working Paper.

Rice, Jennifer K. 2013. "Learning from Experience? Evidence on the Impact and Distribution of Teacher Experience and the Implications for Teacher Policy." *Education Finance and Policy* 8(3): 332–348.

Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data." *American Economic Review* 94(2): 247-252.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy* 6(1): 43-74.

Rothstein, Jesse. (2015) "Teacher Quality Policy When Supply Matters." *American Economic Review* 105(1):100-130.

Schmutte, Ian M. 2015. "Job Referral Networks and the Determination of Earnings in Local Labor Markets" *Journal of Labor Economics,* 33(1)

Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24: 97-117

Stanton, Christopher, and Catherine Thomas. 2016. "Landing the First Job: The Value of Intermediaries in Online Hiring." *Review of Economic Studies* 83(2): 810-854.

Steinberg, Matthew P., and Morgaen L. Donaldson. 2016. "The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era." *Education Finance and Policy 11* (3):340-359.

Taylor, Eric S. 2018. “Skills, Job Tasks, and Productivity in Teaching: Evidence from a Randomized Trial of Instruction Practices.” *Journal of Labor Economics*, *36* (3): 711-742.

Taylor, Eric S. and John H. Tyler. 2012. The Effect of Evaluation on Teacher Performance.” *American Economic Review 107* (7): 3638-3651.

Treu, Rolf M. 2014. Vergara vs. State of California Tentative Decision. Available http://studentsmatter.org/wp-content/uploads/2014/06/Tenative-Decision.pdf. Accessed 7 June 2017.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. 2009. *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Effectiveness.* New York: The New Teacher Project.

White, Mark. 2018. “Accounting for Student Composition in Estimates of Teacher Quality from Classroom Observation Instruments.” Working paper. University of Michigan.

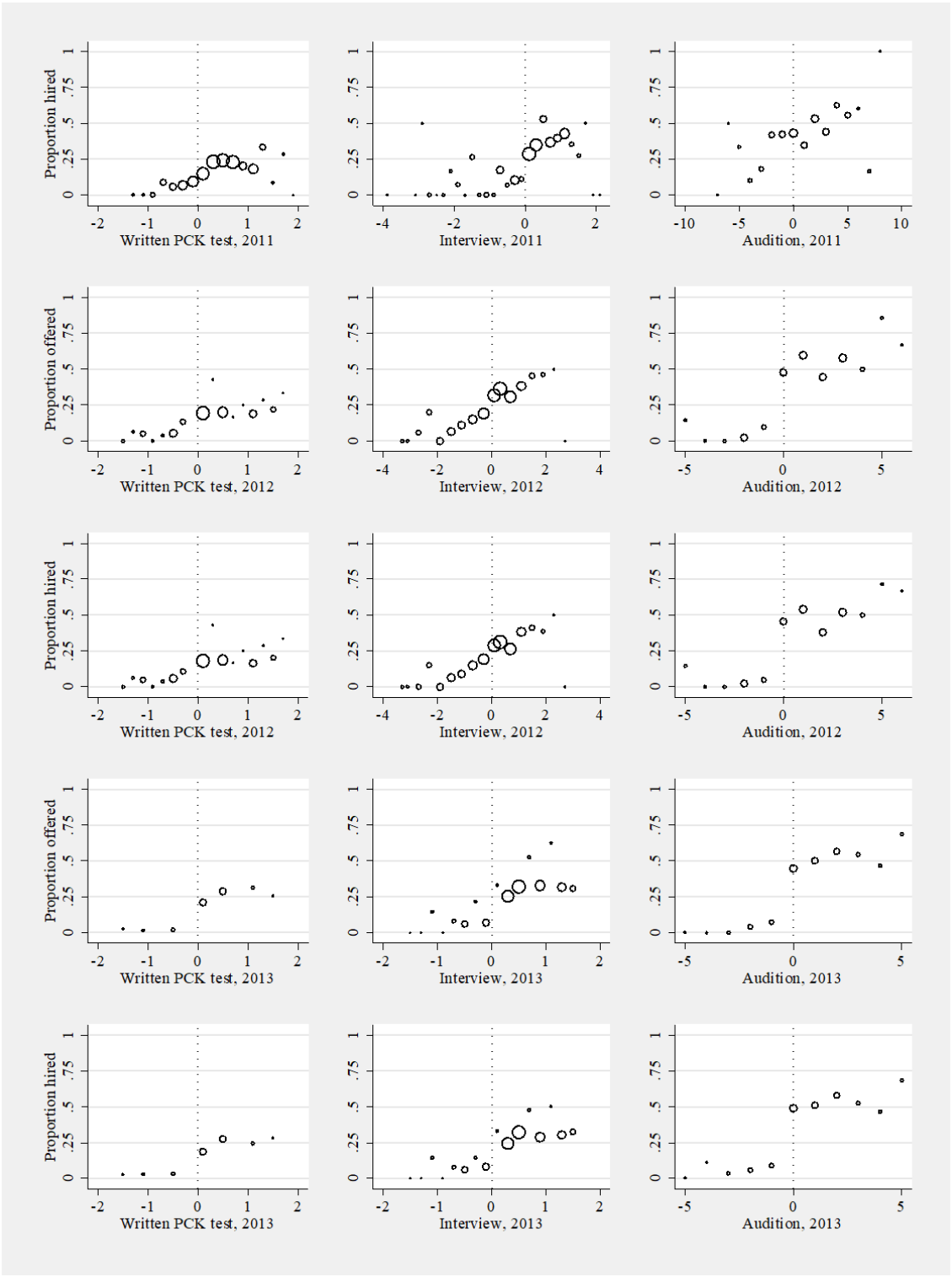



Figure 1—Screening scores and the probability of job offer and hire

Note: Circles indicate the proportion of applicants offered or hired (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicants. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.

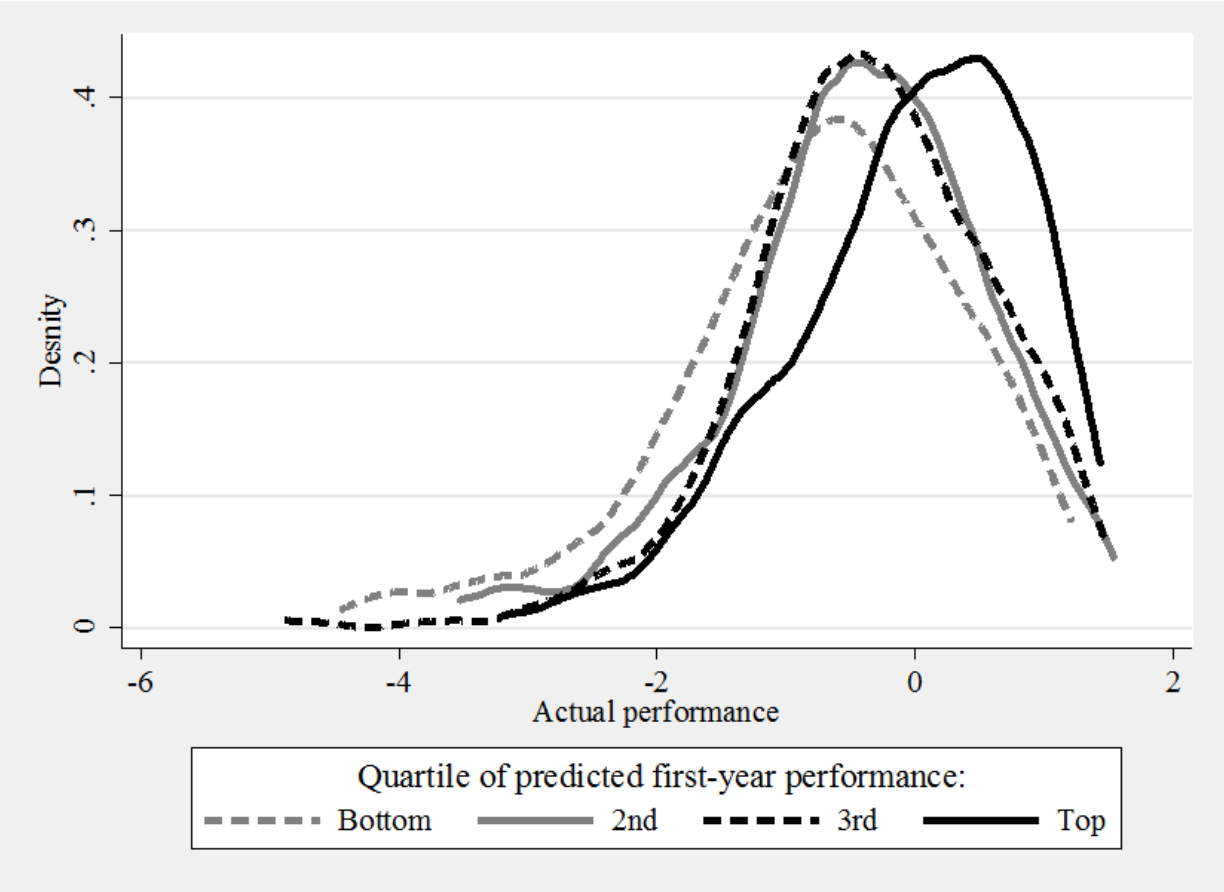



Figure 2—Actual first-year performance and performance predicted by application

Note: Kernel densities of actual first-year performance of hired teachers estimated separately by quartile of predicted first-year performance. Predicted performance is the fitted value obtained after estimating a regression similar to Table 6 Column 1 with two modifications: (i) We use only observations from each new hire's first year on the job. The coefficient estimates on this restricted sample are quite similar to those obtained from the full sample. (ii) The covariates include each individual component of the academic index and screening scores index separately as in Appendix Table 6 Column 2. The specification also includes subject-taught by year fixed effects, but the fitted value (predicted performance) does not include subject-taught by year effects. Last, predicted performance is estimated using a leave-one-out or jackknife procedure: to obtain the prediction for teacher *i*, we estimate our model using all observations except *i*.

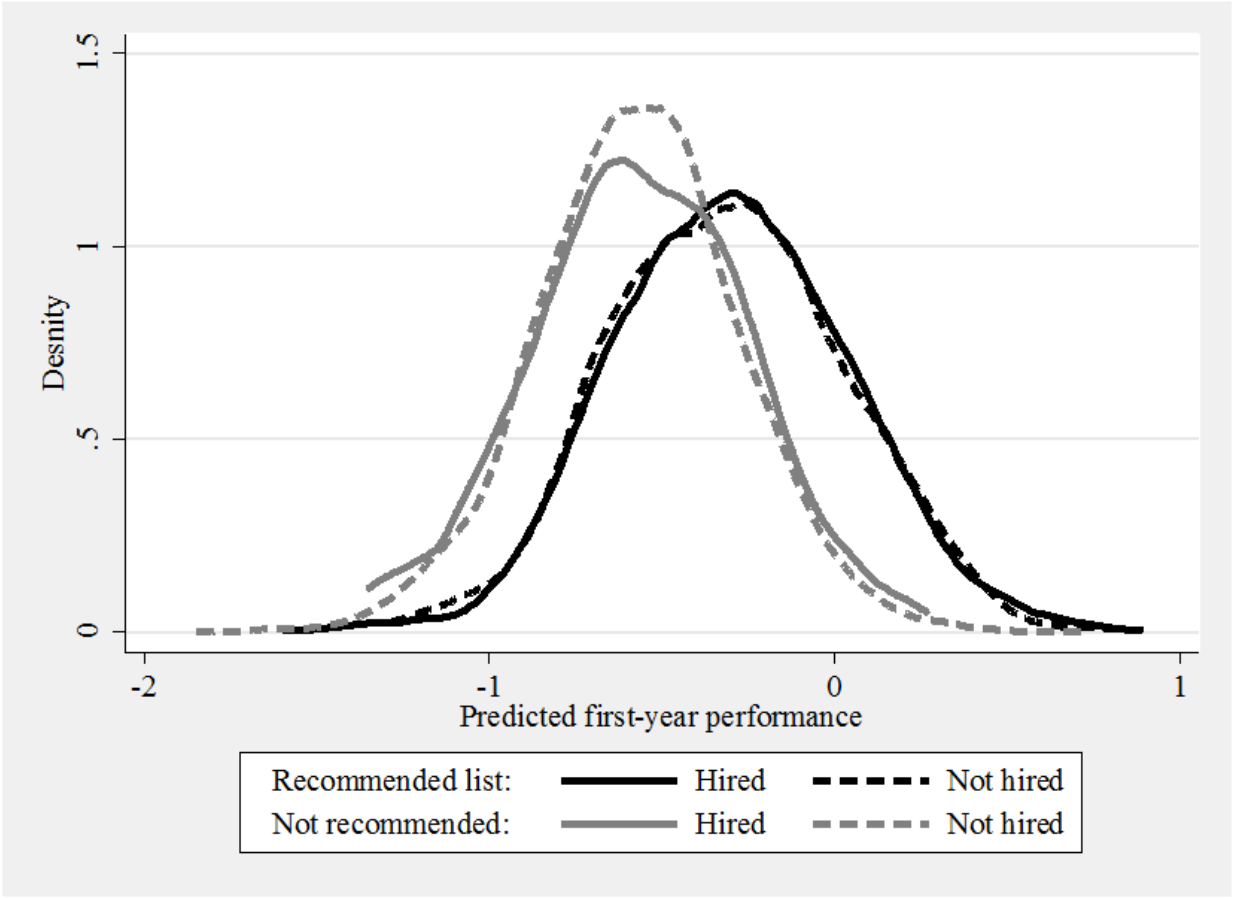



Figure 3—Predicted first-year performance for applicants

Note: Kernel densities of predicted first-year performance estimated separately for applicants in four mutually-exclusive categories: the intersection of applicants recommended or not and applicants hired or not. Predicted performance is obtained as follows: First, using the sample of new hires in their first year at DCPS, fit a regression similar to Table 6 Column 1, except that covariates include each individual component of the academic index and screening scores index separately as in Appendix Table 6 Column 2. The specification also includes subject-taught by year fixed effects. Second, the estimated coefficients from that regression are applied to the applicant sample. This predicted performance measure does not include differences between the subject-taught by year fixed effect groups.

Table 1—Characteristics of TeachDC applicants

| | All applicants | | Applicants hired | |
|---|---|---|---|---|
| | Obs. | Mean (st.dev.) | Obs. | Mean (st.dev.) |
| Hired | 7,442 | 0.13 | 982 | 1 |
| Prior teaching experience | 7,314 | | 978 | |
| Novice | | 0.33 | | 0.28 |
| 1 to 2 | | 0.17 | | 0.19 |
| 3 to 5 | | 0.18 | | 0.20 |
| 6 to 10 | | 0.17 | | 0.20 |
| 11 or more | | 0.14 | | 0.14 |
| Undergraduate GPA | 7,112 | 3.40 | 939 | 3.42 |
| | | (0.43) | | (0.44) |
| SAT math+verbal (or ACT equiv) | 4,600 | 1148.71 | 674 | 1148.75 |
| | | (175.14) | | (168.58) |
| Undergraduate college Barron's ranking | 6,588 | 2.81 | 907 | 2.91 |
| | | (1.24) | | (1.26) |
| Master's degree or higher | 7,442 | 0.51 | 982 | 0.54 |
| Location of undergrad or grad school | 7,076 | | 940 | |
| DC | | 0.12 | | 0.17 |
| Maryland or Virginia | | 0.28 | | 0.28 |
| Outside DC, MD, VA | | 0.60 | | 0.55 |
| Academics index | 7,442 | 0.00 | 982 | 0.07 |
| | | (1.00) | | (1.00) |
| Screening scores index | 4,668 | 0.00 | 799 | 0.50 |
| | | (1.00) | | (0.71) |

Note: Excluding applicants who were not eligible for a teaching license in DC. Location indicators are mutually exclusive, applicants with multiple locations coded based on location nearest DC. “Academics index” is the average of standardized values of GPA, SAT, Barron’s ranking, and MA degree (or the observed subset of those four values), which is then standardized. “Screening scores index” is the sum of standardized PCK, interview, and audition scores (or the observed subset of those three), which is then standardized.

Table 2—Pairwise correlations of applicant characteristics and scores

| | Exper. | DC | MD-VA | GPA | SAT | Barron's | MA | PCK | Interv. | Aud. |
|---|---|---|---|---|---|---|---|---|---|---|
| Years of prior experience | 1 | | | | | | | | | |
| *Location* | | | | | | | | | | |
| Washington DC | -0.01 | 1 | | | | | | | | |
| Maryland or Virginia | 0.02 | -0.23 | 1 | | | | | | | |
| *Academic Measures* | | | | | | | | | | |
| Undergraduate GPA | -0.10 | 0.04 | -0.07 | 1 | | | | | | |
| SAT math+verbal | -0.05 | 0.03 | -0.05 | 0.31 | 1 | | | | | |
| Barron's rank | -0.14 | 0.06 | -0.01 | 0.13 | 0.34 | 1 | | | | |
| Master's degree | 0.16 | 0.00 | 0.03 | 0.08 | 0.07 | 0.08 | 1 | | | |
| *Selection Scores* | | | | | | | | | | |
| PCK written test | -0.11 | 0.01 | -0.04 | 0.16 | 0.22 | 0.19 | 0.04 | 1 | | |
| Interview | -0.02 | 0.02 | -0.02 | 0.12 | 0.13 | 0.08 | 0.03 | 0.10 | 1 | |
| Audition | 0.04 | 0.01 | 0.00 | 0.06 | 0.08 | 0.03 | 0.04 | 0.10 | 0.22 | 1 |

Note: Pairwise correlations of applicant characteristics and scores. Maximum observations for a cell is 7,442.

Table 3—Hiring

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Undergrad GPA (std) | 0.006 | -0.014** | -0.011* | | |
| | (0.004) | (0.004) | (0.004) | | |
| SAT/ACT math+verbal (std) | 0.000 | -0.016** | -0.015** | | |
| | (0.005) | (0.005) | (0.005) | | |
| Barron's Rank (linear 0-5) | 0.009* | -0.002 | -0.000 | | |
| | (0.004) | (0.003) | (0.004) | | |
| Master's degree or higher | 0.009 | -0.001 | -0.007 | | |
| | (0.008) | (0.007) | (0.008) | | |
| Academics index | 0.005 | -0.012** | | -0.015** | -0.015** |
| | (0.004) | (0.004) | | (0.004) | (0.004) |
| PCK written test (std) | 0.060** | 0.008+ | 0.008 | | |
| | (0.005) | (0.005) | (0.006) | | |
| Interview (std) | 0.108** | 0.028** | 0.024** | | |
| | (0.008) | (0.009) | (0.009) | | |
| Audition (std) | 0.158** | 0.053** | 0.050** | | |
| | (0.013) | (0.018) | (0.018) | | |
| Screening scores index | 0.069** | 0.023** | | 0.071** | 0.025** |
| | (0.006) | (0.006) | | (0.006) | (0.006) |
| Years prior experience | | | | | |
| 1 to 2 | 0.029* | 0.023* | 0.025* | 0.031** | 0.026* |
| | (0.012) | (0.011) | (0.011) | (0.011) | (0.011) |
| 3 to 5 | 0.027* | 0.024* | 0.025* | 0.033** | 0.028** |
| | (0.012) | (0.010) | (0.010) | (0.011) | (0.010) |
| 6 to 10 | 0.040** | 0.042** | 0.042** | 0.050** | 0.046** |
| | (0.012) | (0.011) | (0.011) | (0.011) | (0.011) |
| 11 or more | 0.006 | 0.026* | 0.029* | 0.036** | 0.033** |
| | (0.012) | (0.011) | (0.012) | (0.012) | (0.012) |
| Location of undergrad or grad school | | | | | |
| DC | 0.056** | 0.056** | 0.057** | 0.058** | 0.058** |
| | (0.014) | (0.013) | (0.013) | (0.013) | (0.013) |
| Maryland or Virginia | 0.013 | 0.026** | 0.025** | 0.024** | 0.026** |
| | (0.009) | (0.008) | (0.008) | (0.009) | (0.008) |
| Recommended-pool by year FE | | √ | √ | | √ |
| Adjusted R-squared | | | 0.207 | 0.151 | 0.205 |
| F-statistic subject-applied by year FE | | | 1.26 | 1.47 | 1.25 |
| p-value | | | 0.053 | 0.003 | 0.057 |
| F-statistic recommended-pool by year FE | | | 47.5 | | 75.3 |
| p-value | | | 0.000 | | 0.000 |

Note: Estimates from linear regressions with 7,442 observations, where an indicator for being hired is the dependent variable. In columns 1-2 each group of coefficients separated by a solid line are estimates from a separate regression. Columns 3-5 each report estimates from a single regression. Each specification includes year-by-subject-applied fixed effects. Location indicators are mutually exclusive, applicants with multiple locations coded based on location nearest DC. The recommended-pool by year FE include four mutually exclusive indicators: (i)-(iii) applicants who pass the final audition stage in 2011, 2012, and 2013 respectively, and (iv) applicants in 2011 who pass the interview but not the audition (see the text for more details on this fourth category). The left-out category is all other applicants. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate.

+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 4—Offers and hiring

| | 2012-2013 | | | | 2011-13 | |
|---|---|---|---|---|---|---|
| | Offered | | Hired | | Hired | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Academics index | -0.010* | -0.011* | -0.012** | -0.013** | -0.015** | -0.015** |
| | (0.005) | (0.005) | (0.005) | (0.004) | (0.004) | (0.004) |
| Screening scores index | 0.091** | 0.036** | 0.077** | 0.023** | 0.071** | 0.025** |
| | (0.008) | (0.008) | (0.007) | (0.008) | (0.006) | (0.006) |
| Years prior experience | | | | | | |
| 1 to 2 | 0.032* | 0.028* | 0.027* | 0.023+ | 0.031** | 0.026* |
| | (0.014) | (0.014) | (0.013) | (0.013) | (0.011) | (0.011) |
| 3 to 5 | 0.079** | 0.074** | 0.031* | 0.026* | 0.033** | 0.028** |
| | (0.015) | (0.014) | (0.013) | (0.012) | (0.011) | (0.010) |
| 6 to 10 | 0.081** | 0.077** | 0.054** | 0.049** | 0.050** | 0.046** |
| | (0.015) | (0.014) | (0.013) | (0.013) | (0.011) | (0.011) |
| 11 or more | 0.078** | 0.077** | 0.042** | 0.040** | 0.036** | 0.033** |
| | (0.016) | (0.016) | (0.014) | (0.014) | (0.012) | (0.012) |
| Location of undergrad or grad school | | | | | | |
| DC | 0.062** | 0.061** | 0.046** | 0.045** | 0.058** | 0.058** |
| | (0.016) | (0.016) | (0.014) | (0.014) | (0.013) | (0.013) |
| Maryland or Virginia | 0.028* | 0.030** | 0.025* | 0.028** | 0.024** | 0.026** |
| | (0.012) | (0.011) | (0.010) | (0.010) | (0.009) | (0.008) |
| Recommended-pool by year FE | | √ | | | | √ |
| Adjusted R-squared | 0.149 | 0.203 | 0.169 | 0.238 | 0.151 | 0.205 |
| F-statistic subject-applied by year FE | 1.92 | 1.79 | 1.48 | 1.29 | 1.47 | 1.25 |
| p-value | 0.000 | 0.000 | 0.011 | 0.074 | 0.003 | 0.057 |
| F-statistic recommended-pool by year FE | | 116.3 | | 126.4 | | 75.3 |
| p-value | | 0.000 | | 0.000 | | 0.000 |

Note: Estimates from linear regressions with 5,082 observations in columns 1-4, and 7,442 in columns 5-6, where an indicator for being hired or offered is the dependent variable as indicated in the column headings. Columns 5-6 reproduce columns 4-5 in Table 3 for convenient comparison. The specification for columns 1-4 is the same as columns 5-6, except for the sample and dependent variable. Each specification includes year-by-subject-applied fixed effects. For additional details see the note in Table 3.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 5—School achievement and characteristics of hires and offers

| | Dep. var. = proportion of students passing state test at the hiring (offering) school | | | |
|---|---|---|---|---|
| | Hires | | Offers | |
| | (1) | (2) | (3) | (4) |
| Academics index | 0.041** | 0.041** | 0.033** | 0.033** |
| | (0.011) | (0.011) | (0.009) | (0.010) |
| Screening scores index | -0.018 | -0.018 | -0.013 | -0.012 |
| | (0.012) | (0.012) | (0.011) | (0.011) |
| Years prior experience | | | | |
| 1 to 2 | 0.028 | 0.028 | -0.015 | -0.016 |
| | (0.025) | (0.026) | (0.026) | (0.026) |
| 3 to 5 | 0.002 | 0.001 | -0.014 | -0.013 |
| | (0.025) | (0.026) | (0.022) | (0.021) |
| 6 to 10 | -0.023 | -0.023 | -0.043+ | -0.043+ |
| | (0.026) | (0.027) | (0.024) | (0.025) |
| 11 or more | -0.031 | -0.031 | -0.048 | -0.048 |
| | (0.030) | (0.030) | (0.032) | (0.032) |
| Location of undergrad or grad school | | | | |
| DC | 0.001 | 0.001 | -0.009 | -0.010 |
| | (0.022) | (0.022) | (0.021) | (0.022) |
| Maryland or Virginia | -0.029 | -0.029 | -0.028 | -0.029 |
| | (0.021) | (0.021) | (0.020) | (0.020) |
| Recommended-pool by year FE | | √ | | √ |
| Subject-taught by year FE | √ | √ | | |
| Subject-applied by year FE | | | √ | √ |
| Adjusted R-squared | 0.065 | 0.062 | 0.032 | 0.029 |
| F-statistic subject-applied (-taught) by year FE | 1.61 | 1.60 | 1.31 | 1.30 |
| p-value | 0.016 | 0.017 | 0.111 | 0.112 |
| F-statistic recommended-pool by year FE | | 0.0 | | 0.2 |
| p-value | | 0.961 | | 0.826 |

Note: Estimates from separate linear regressions with 575 observations in Columns 1-2, and 753 in Columns 3-4. The estimation sample only includes applicants who were hired in 2012 or 2013 in Columns 1-2, or applicants who were offered jobs in 2012 or 2013 in Columns 3-4. The dependent variable is the proportion of students in the hiring (offering) school who passed (scored proficient or advanced) on the 2010-11 DC-CAS. Location indicators are mutually exclusive, applicants with multiple locations coded based on location nearest DC. The recommended-pool by year FE include four mutually exclusive indicators: (i)-(iii) applicants who pass the final audition stage in 2011, 2012, and 2013 respectively, and (iv) applicants in 2011 who pass the interview but not the audition (see the text for more details on this fourth category). The left-out category is all other applicants. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate. Clustered (school) standard errors in parentheses.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 6—Job performance

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Academics index | 0.210** | 0.211** | 0.148** | 0.176** |
| | (0.026) | (0.027) | (0.041) | (0.025) |
| Screening scores index | 0.212** | 0.220** | 0.168** | 0.172** |
| | (0.030) | (0.031) | (0.041) | (0.029) |
| Years prior experience | | | | |
| 1 to 2 | 0.061 | 0.059 | 0.069 | 0.023 |
| | (0.070) | (0.070) | (0.098) | (0.064) |
| 3 to 5 | 0.067 | 0.060 | 0.136 | 0.092 |
| | (0.070) | (0.071) | (0.103) | (0.065) |
| 6 to 10 | -0.048 | -0.050 | 0.029 | 0.024 |
| | (0.070) | (0.071) | (0.101) | (0.064) |
| 11 or more | -0.116 | -0.120 | -0.104 | -0.073 |
| | (0.093) | (0.093) | (0.152) | (0.085) |
| Location of undergrad or grad school | | | | |
| DC | -0.055 | -0.062 | -0.036 | -0.003 |
| | (0.069) | (0.070) | (0.101) | (0.065) |
| Maryland or Virginia | -0.084 | -0.082 | -0.096 | -0.000 |
| | (0.059) | (0.059) | (0.076) | (0.053) |
| Predicted probability of hire | | | -0.659 | |
| | | | (0.830) | |
| Predicted probability of hire ^ 2 | | | 0.728 | |
| | | | (0.865) | |
| Recommended-pool by year FE | | √ | | √ |
| School FE | | | | √ |
| Adjusted R-squared | 0.193 | 0.194 | 0.223 | 0.351 |
| F-statistic recommended-pool by year FE | | 0.829 | | 1.059 |
| p-value | | 0.547 | | 0.386 |
| F-statistic school FE | | | | 12.845 |
| p-value | | | | 0.000 |
| F-statistic predicted pr. hire terms jointly zero | | | 0.350 | |
| p-value | | | 0.702 | |
| F-statistic excluded instruments | | | 42.41 | |

Note: Estimates from least squares regressions with 3,015 teacher-by-year observations from school years 2011-12 through 2016-17, and 927 unique teachers. The dependent variable is job performance measured by the first predicted factor of IMPACT evaluation component scores, standardized. Location indicators are mutually exclusive, applicants with multiple locations coded based on location nearest DC. The recommended-pool by year FE include four mutually exclusive indicators: (i)-(iii) applicants who pass the final audition stage in 2011, 2012, and 2013 respectively, and (iv) applicants in 2011 who pass the interview but not the audition (see the text for more details on this fourth category). The left-out category is all other applicants. All specifications include year-by-subject-taught fixed effects, and fixed effects for the number of years since hire. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate. Clustered (teacher) standard errors in parentheses.

For Column 3 the predicted probability of hire is estimated in an auxiliary regression using the specification in Table 3 Column 3 but without the recommended pool fixed effects and with additional instruments added as regressors. The instruments are two indicator variables: (i) Applicants in any year who scored above the stage 4 cut-score designated as the threshold for the recommended pool. (ii) Applicants in 2011 who scored above the stage 3 cut-score (see the text for more details on this category).

+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 7—IMPACT score components

| | | Component scores | | | | |
|---|---|---|---|---|---|---|
| | | Class observation | | | | |
| | IMPACT score | Principal | Master educator | CP | CSC | TAS |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Academics index | 0.172** | 0.140** | 0.129** | 0.045** | 0.189** | 0.125** |
| | (0.026) | (0.028) | (0.026) | (0.008) | (0.026) | (0.025) |
| Screening scores index | 0.175** | 0.200** | 0.150** | 0.034** | 0.195** | 0.123** |
| | (0.030) | (0.033) | (0.029) | (0.010) | (0.031) | (0.030) |
| Years prior experience | | | | | | |
| 1 to 2 | 0.077 | 0.129+ | 0.145* | -0.004 | -0.014 | 0.049 |
| | (0.068) | (0.071) | (0.066) | (0.021) | (0.070) | (0.068) |
| 3 to 5 | 0.135* | 0.196** | 0.182** | -0.013 | -0.032 | 0.030 |
| | (0.069) | (0.071) | (0.067) | (0.023) | (0.068) | (0.069) |
| 6 to 10 | 0.020 | 0.082 | 0.034 | -0.059* | -0.096 | 0.019 |
| | (0.071) | (0.073) | (0.075) | (0.024) | (0.068) | (0.069) |
| 11 or more | -0.047 | -0.040 | -0.034 | -0.023 | -0.171+ | -0.119 |
| | (0.089) | (0.095) | (0.085) | (0.028) | (0.090) | (0.089) |
| Location of undergrad or grad school | | | | | | |
| DC | -0.001 | 0.060 | -0.022 | -0.029 | -0.112+ | 0.011 |
| | (0.068) | (0.070) | (0.069) | (0.023) | (0.067) | (0.066) |
| Maryland or Virginia | -0.040 | -0.021 | -0.016 | -0.028 | -0.105+ | -0.004 |
| | (0.057) | (0.060) | (0.057) | (0.019) | (0.057) | (0.055) |
| Recommended-pool by year FE | √ | √ | √ | √ | √ | √ |
| F-statistic rec.-pool by year FE | 0.7 | 0.5 | 0.8 | 1.4 | 1.0 | 2.0 |
| p-value | 0.636 | 0.816 | 0.598 | 0.211 | 0.421 | 0.064 |

Note: The details of estimation are identical to Table 6 Column 2, except that each regression reported above has a different dependent variable as described in the column headers.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 8—Job performance by subject and grade taught

| | Elementary and ECE | MS/HS core | Special education | Other specialties |
|---|---|---|---|---|
| | (1) | (2) | (3) | (3) |
| Academics index | 0.199** | 0.211** | 0.224** | 0.220** |
| | (0.045) | (0.050) | (0.079) | (0.061) |
| Screening scores index | 0.236** | 0.191** | 0.320** | 0.289** |
| | (0.056) | (0.060) | (0.081) | (0.079) |
| Years prior experience | | | | |
| 1 to 2 | 0.054 | -0.047 | -0.011 | 0.227 |
| | (0.099) | (0.172) | (0.227) | (0.172) |
| 3 to 5 | 0.036 | 0.232+ | -0.032 | -0.015 |
| | (0.131) | (0.132) | (0.171) | (0.170) |
| 6 to 10 | -0.170 | -0.061 | -0.210 | 0.275+ |
| | (0.113) | (0.140) | (0.192) | (0.158) |
| 11 or more | -0.216 | -0.110 | -0.160 | -0.084 |
| | (0.183) | (0.192) | (0.187) | (0.214) |
| Location of undergrad or grad school | | | | |
| DC | -0.090 | 0.035 | -0.206 | -0.133 |
| | (0.112) | (0.134) | (0.196) | (0.176) |
| Maryland or Virginia | -0.091 | -0.072 | 0.002 | -0.230 |
| | (0.100) | (0.126) | (0.141) | (0.143) |
| Recommended-pool by year FE | √ | √ | √ | √ |
| Adjusted R-squared | 0.170 | 0.200 | 0.196 | 0.241 |
| F-statistic recommended-pool by year FE | 1.8 | 0.5 | 1.3 | 1.4 |
| p-value | 0.089 | 0.813 | 0.282 | 0.220 |
| Number of observations | 1127 | 764 | 478 | 633 |

Note: The details of estimation are identical to Table 6 Column 2, except that each regression above is estimated with a subsample defined in the column headers. "Middle/High school core subjects" includes English, math, sciences, and social studies. "Special education" includes teachers specializing in special education at any grade level. "Other specialties" includes arts, foreign languages, physical education and health, English language learners at any grade level, and others.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Table 9—Job performance for teachers with individual value-added scores

| | Overall performance | | Value-added score (math and reading observations pooled) | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Academics index | 0.219** | 0.138 | 0.186* | 0.026 |
| | (0.081) | (0.100) | (0.084) | (0.116) |
| Screening scores index | 0.213** | 0.083 | -0.112 | 0.097 |
| | (0.076) | (0.088) | (0.085) | (0.121) |
| Years prior experience | | | | |
| 1 to 2 | 0.145 | 0.218 | 0.086 | 0.056 |
| | (0.202) | (0.211) | (0.236) | (0.276) |
| 3 to 5 | 0.372+ | 0.594** | 0.407+ | 0.427 |
| | (0.207) | (0.196) | (0.233) | (0.273) |
| 6 to 10 | 0.092 | 0.194 | 0.269 | 0.070 |
| | (0.187) | (0.189) | (0.202) | (0.260) |
| 11 or more | 0.012 | 0.362 | 0.530* | 0.446 |
| | (0.241) | (0.220) | (0.224) | (0.334) |
| Location of undergrad or grad school | | | | |
| DC | -0.238 | 0.238 | -0.106 | 0.293 |
| | (0.183) | (0.235) | (0.199) | (0.218) |
| Maryland or Virginia | 0.003 | 0.156 | 0.046 | -0.115 |
| | (0.154) | (0.197) | (0.164) | (0.210) |
| Recommended-pool by year FE | √ | √ | √ | √ |
| Subject-by-grade-by-year FE | √ | √ | √ | √ |
| School FE | | √ | | √ |
| Teacher-year-subject observations | | | 433 | 433 |
| Teacher-year observations | 346 | 346 | 346 | 346 |
| Teacher observations | 218 | 218 | 218 | 218 |

Note: Estimates from least squares regressions using data from four school years: 2011-12 to 2013-14, and 2016-17. DCPS did not use value-added scores in 2014-15 and 2015-16. The dependent variables are job performance in Columns 1-2, the same outcome variable as in Table 6; and value-added in either math or reading in Columns 3-4. Both job performance and value-added are standardized (mean zero, standard deviation one) at the teacher level. In Column 1-2 observations are teacher-by-year. In Columns 3-4 observations are teacher-by-year-by-subject, because some teachers have both math and reading value-added scores. All regressions include subject-by-grade-by-year fixed effects, and fixed effects for the number of years since hire. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate. Clustered (teacher) standard errors in parentheses.

+ indicates $p < 0.10$, * 0.05, and ** 0.01

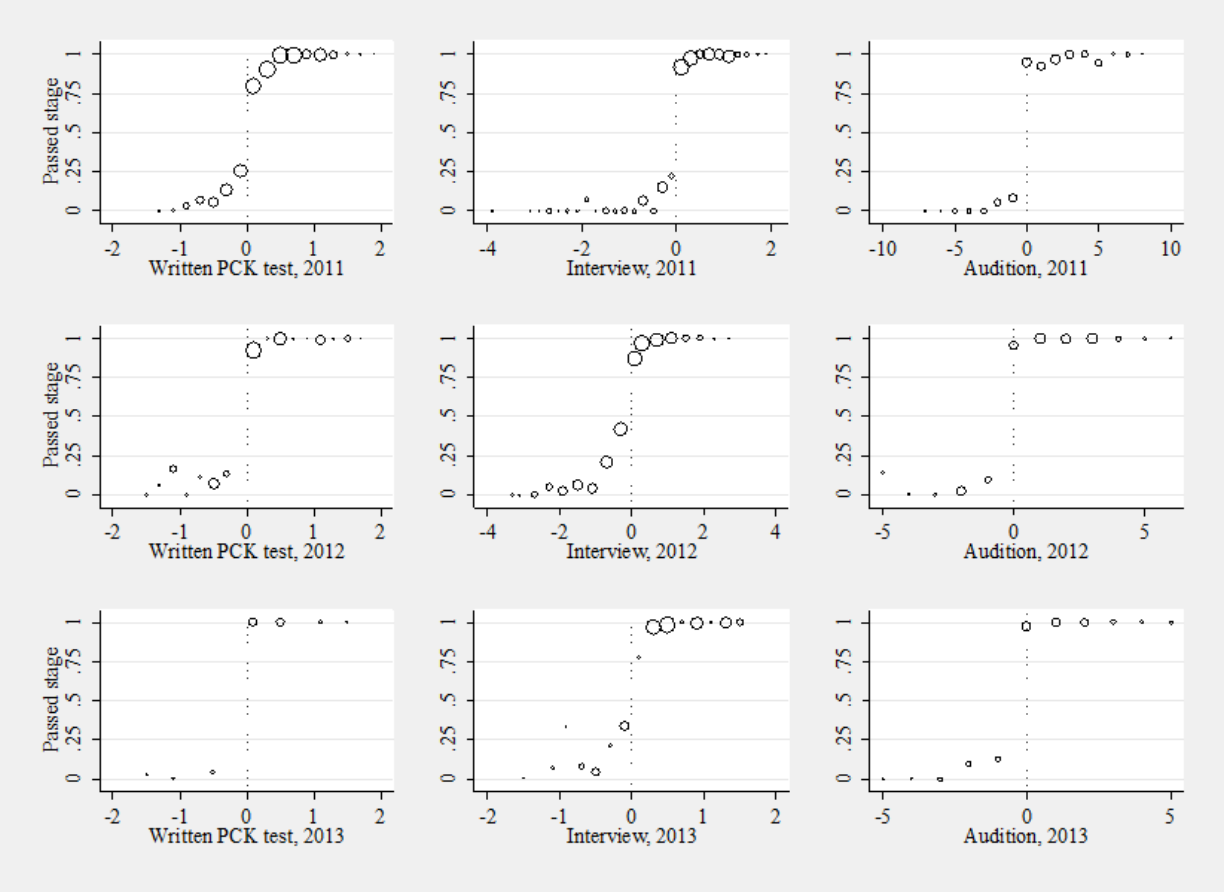
Table 10—Attrition

| | Leave DCPS by the end of… | | | | Leave hiring school by the end of… | | | |
|---|---|---|---|---|---|---|---|---|
| | year 1 | year 2 | year 3 | year 4 | year 1 | year 2 | year 3 | year 4 |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| Academics index | 0.039** | 0.020 | 0.024 | 0.036* | 0.029+ | -0.023 | -0.010 | -0.005 |
| | (0.015) | (0.017) | (0.018) | (0.018) | (0.016) | (0.018) | (0.017) | (0.016) |
| Screening scores index | -0.022 | -0.025 | -0.035 | -0.024 | -0.044* | -0.024 | -0.014 | -0.002 |
| | (0.017) | (0.021) | (0.022) | (0.022) | (0.020) | (0.023) | (0.022) | (0.021) |
| Years prior experience (novice omitted) | | | | | | | | |
| 1 to 2 | 0.014 | 0.015 | 0.015 | 0.016 | 0.008 | 0.022 | 0.001 | 0.043 |
| | (0.043) | (0.050) | (0.051) | (0.051) | (0.047) | (0.052) | (0.049) | (0.046) |
| 3 to 5 | -0.023 | -0.061 | -0.020 | -0.008 | 0.017 | 0.018 | 0.008 | 0.055 |
| | (0.040) | (0.049) | (0.051) | (0.051) | (0.046) | (0.052) | (0.049) | (0.045) |
| 6 to 10 | 0.005 | -0.008 | 0.005 | 0.063 | 0.040 | 0.037 | 0.090+ | 0.095* |
| | (0.042) | (0.049) | (0.050) | (0.050) | (0.047) | (0.051) | (0.046) | (0.044) |
| 11 or more | -0.018 | 0.016 | 0.025 | 0.033 | -0.005 | 0.048 | 0.066 | 0.056 |
| | (0.048) | (0.058) | (0.061) | (0.059) | (0.055) | (0.060) | (0.056) | (0.053) |
| Location of undergrad or grad school | | | | | | | | |
| DC | -0.147** | -0.150** | -0.246** | -0.216** | -0.204** | -0.166** | -0.174** | -0.157** |
| | (0.036) | (0.046) | (0.049) | (0.050) | (0.040) | (0.050) | (0.050) | (0.048) |
| Maryland or Virginia | -0.056+ | -0.053 | -0.080* | -0.082* | -0.055 | -0.041 | -0.022 | -0.012 |
| | (0.033) | (0.039) | (0.040) | (0.040) | (0.038) | (0.041) | (0.038) | (0.035) |
| Adjusted R-squared | 0.030 | 0.006 | 0.044 | 0.030 | 0.033 | 0.012 | 0.033 | 0.024 |
| Dependent variable mean | 0.205 | 0.349 | 0.514 | 0.565 | 0.297 | 0.499 | 0.664 | 0.733 |

Note: Estimates from least squares regressions with 902 teacher observations. The dependent variable is an indicator described in the column headings. Each specification includes year-by-subject-taught fixed effects. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate. + indicates $p < 0.10$, * 0.05, and ** 0.01

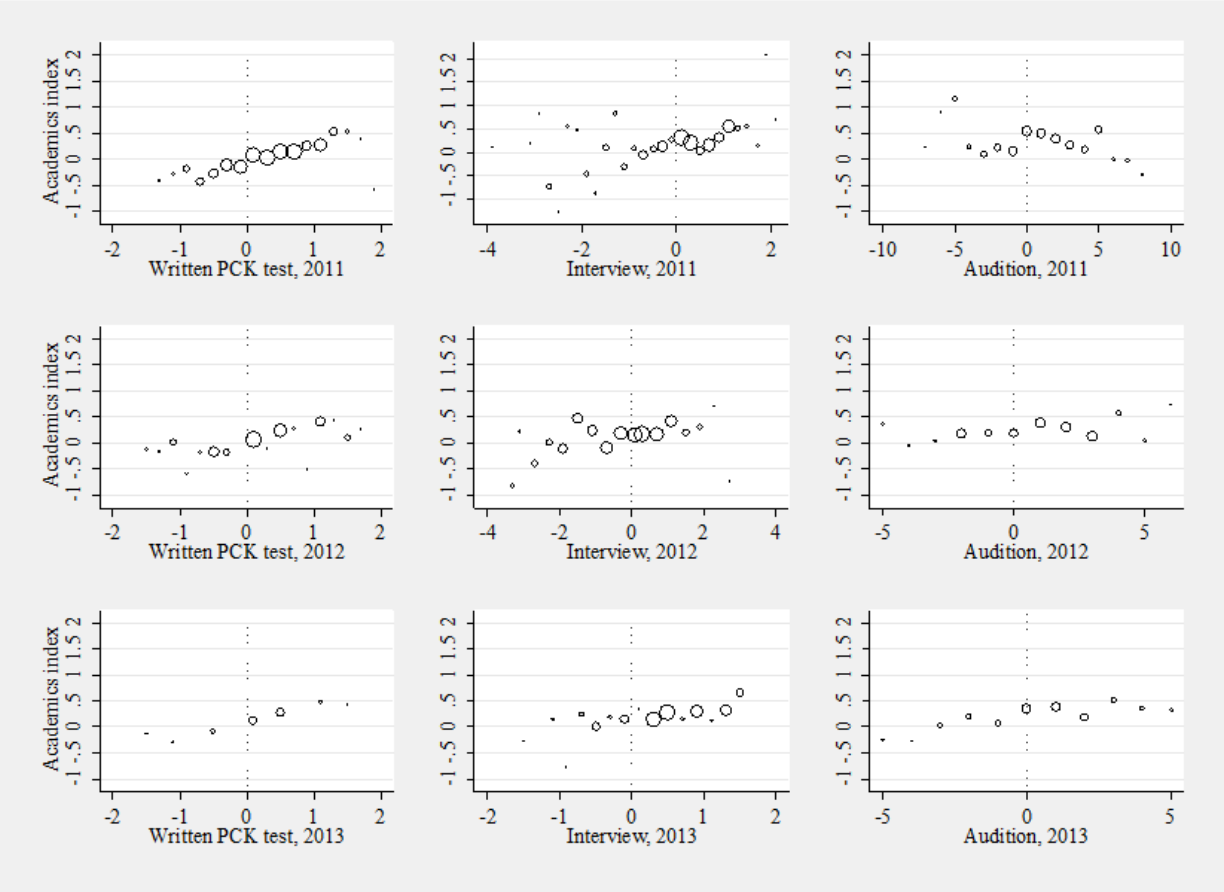
## Online Appendices

## Appendix A: Additional figures and tables




Appendix Figure 1—Screening scores and the probability of passed to the next TeachDC stage

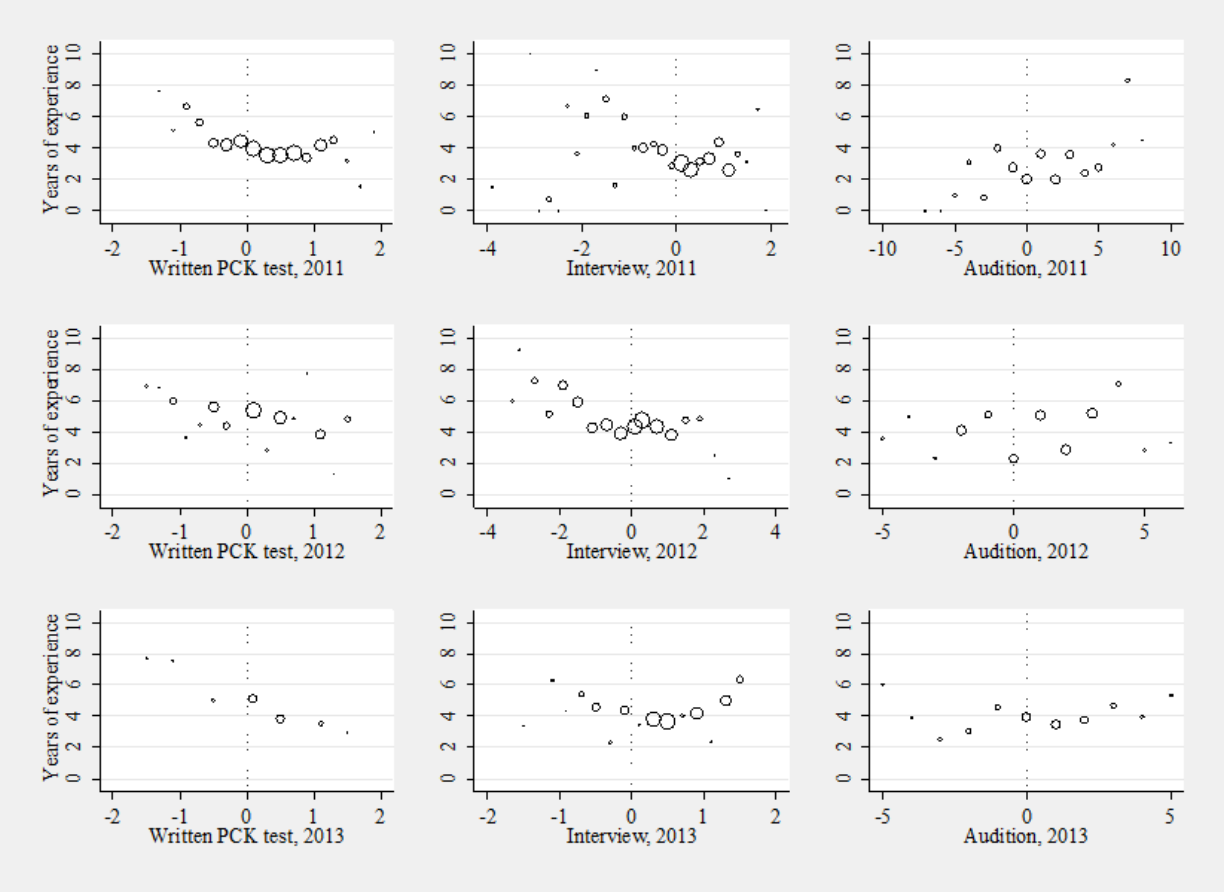
Note: Circles indicate the proportion of applicants invited to participate in the next TeachDC stage (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.




Appendix Figure 2—Screening scores and academic index scores

Note: Circles indicate the mean academic index score of applicants (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.

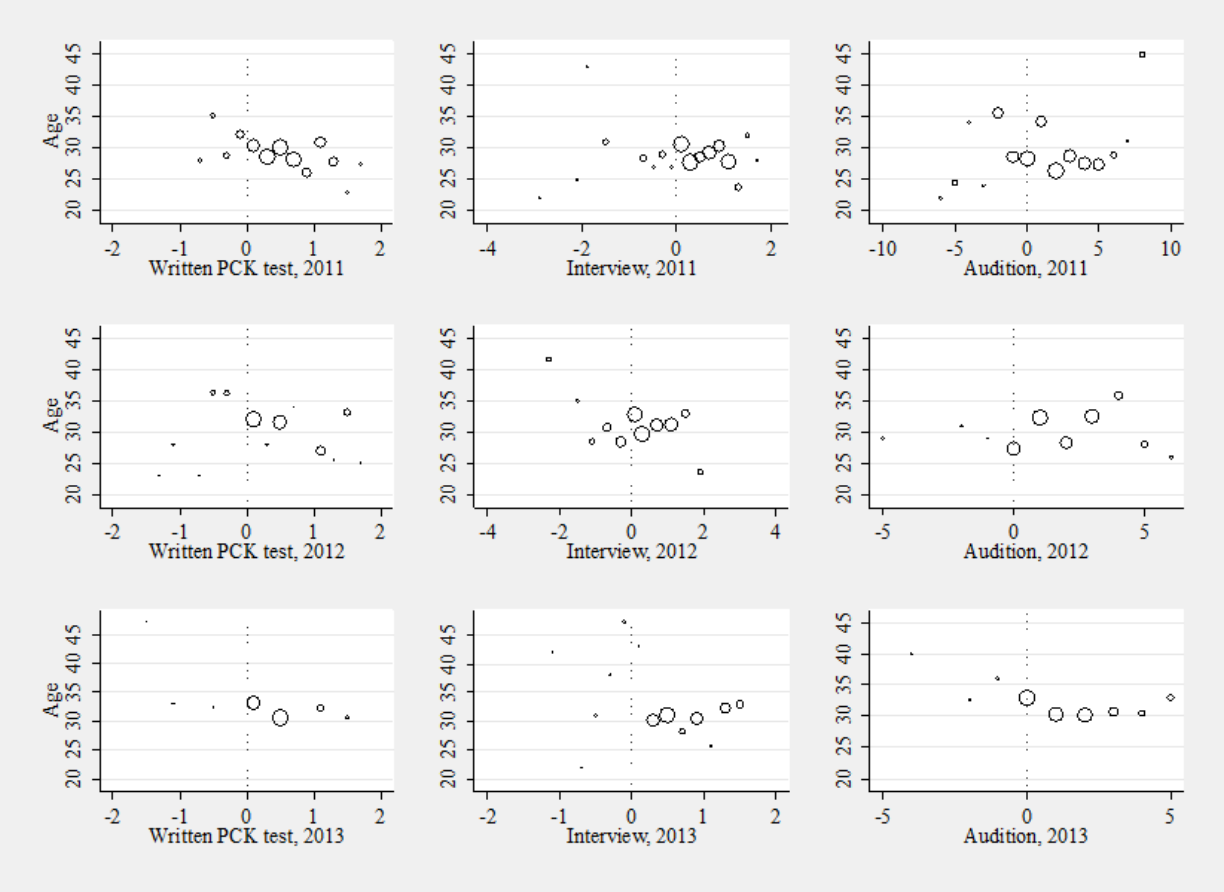
Panel A—Years of teaching experience




Appendix Figure 3—Screening scores and pre-screening applicant characteristics

Note: Circles indicate the mean characteristic (experience, age, attended college in DC) of applicants (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.

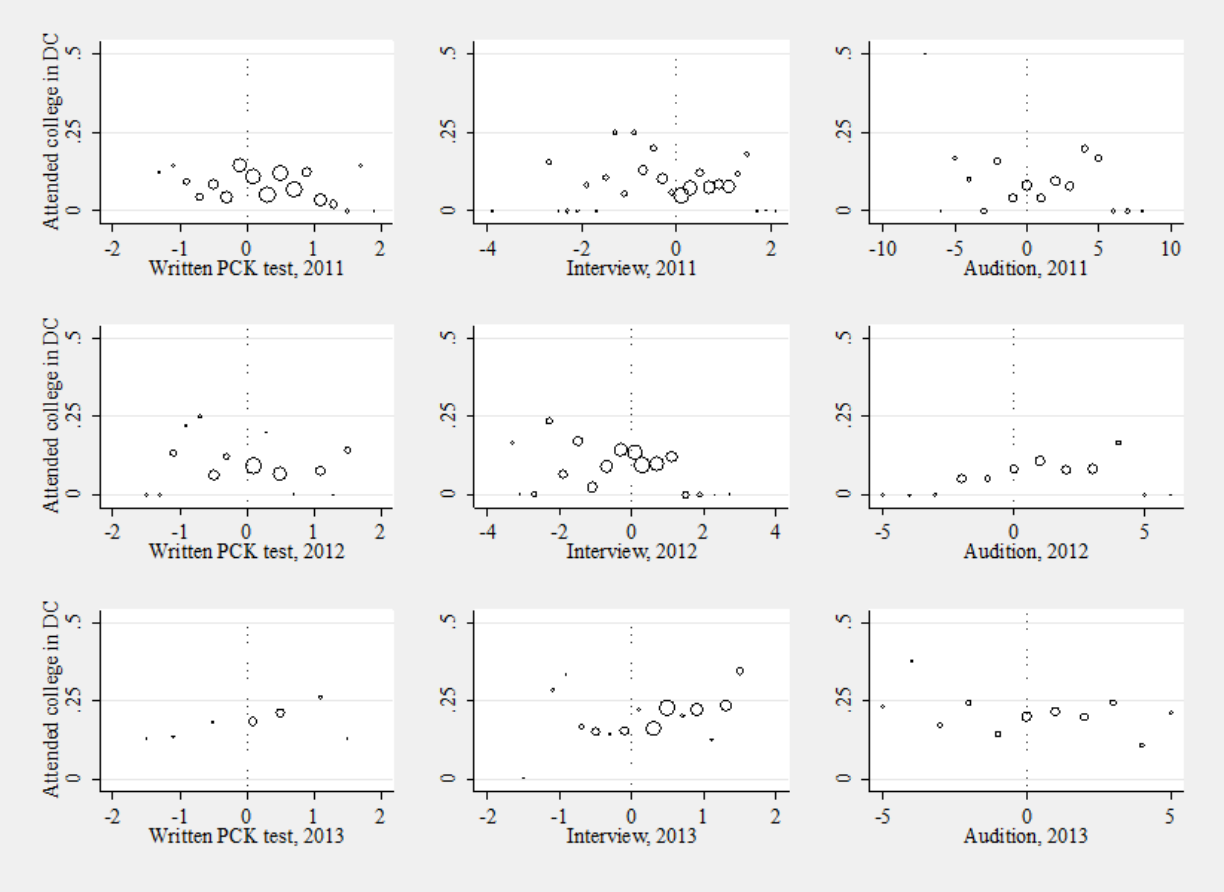
Panel B—Applicant age




Appendix Figure 3—Screening scores and pre-screening applicant characteristics

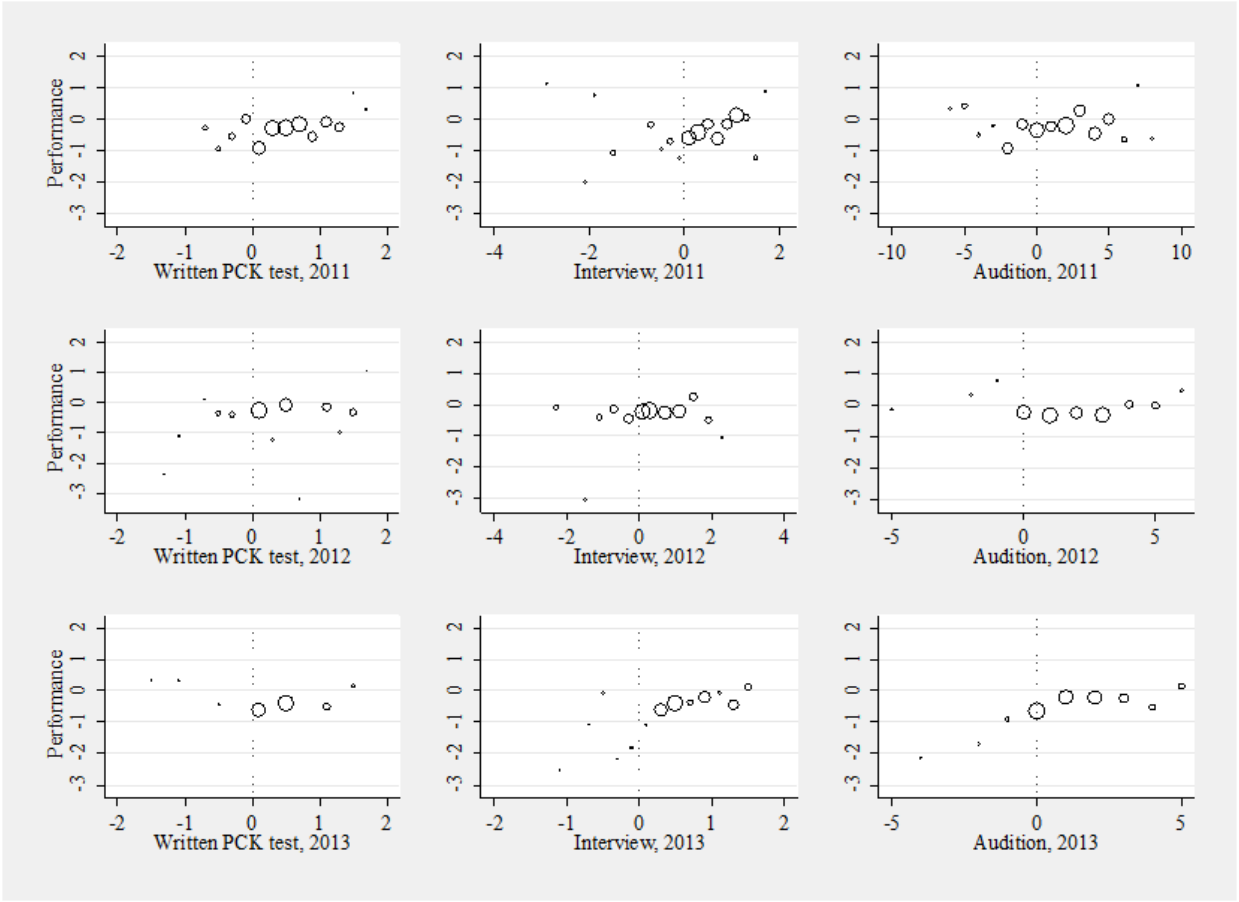
Note: Circles indicate the mean characteristic (experience, age, attended college in DC) of applicants (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.

Panel C—Attended college in DC



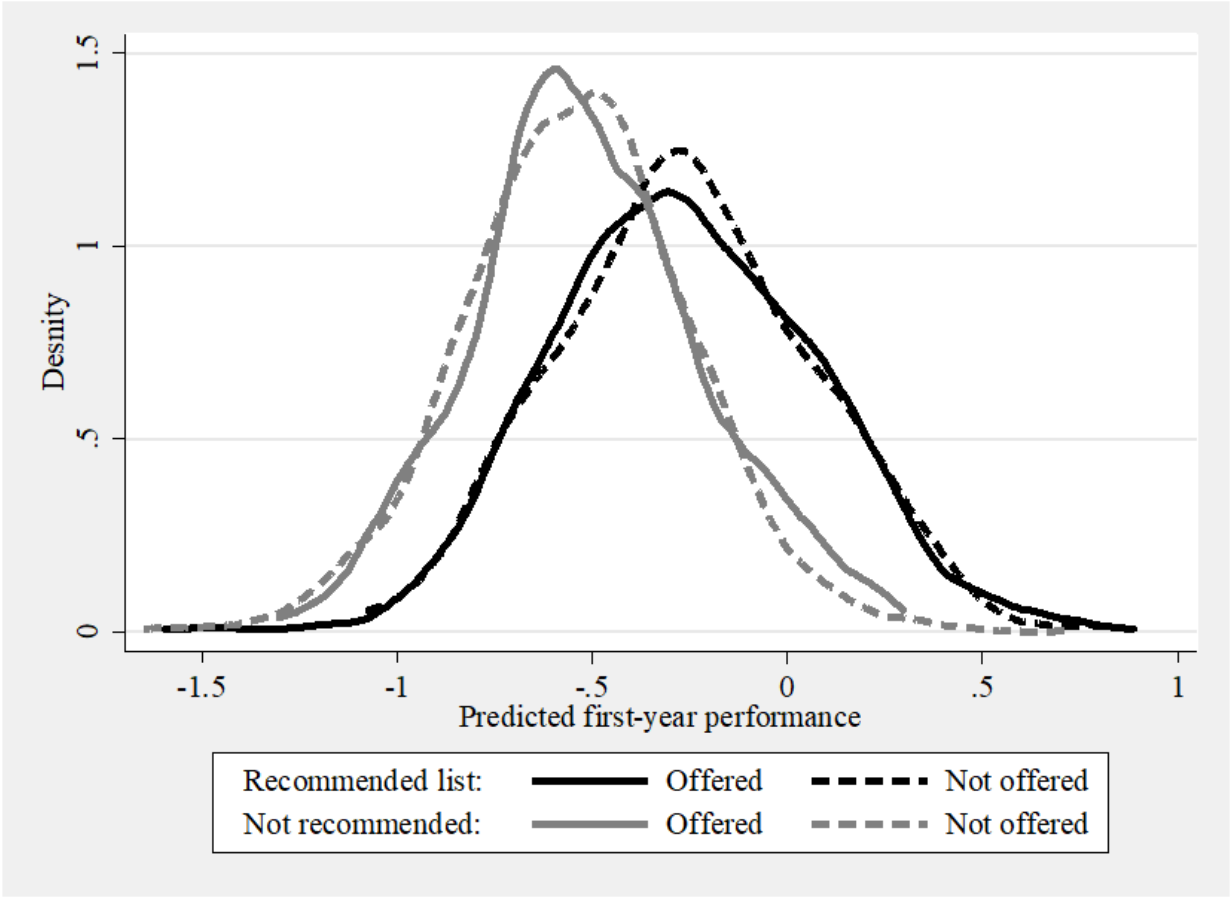

Appendix Figure 3—Screening scores and pre-screening applicant characteristics

Note: Circles indicate the mean characteristic (experience, age, attended college in DC) of applicants (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.




Appendix Figure 4—Screening scores and first-year job performance

Note: Circles indicate the mean first-year actual job performance of hires (y-axis) within bins by screening score (x-axis). Circle markers are scaled by the number of applicant observations in the bin. For PCK and interview the bins are 0.2 points wide. For audition the bins are 1 point wide. Each plot excludes the top and bottom one percent of the x-axis distribution.




Appendix Figure 5—Predicted first-year performance for applicants

Note: Kernel densities of predicted first-year performance estimated separately for applicants in four mutually-exclusive categories: the intersection of applicants recommended or not and applicants offered a job or not. Predicted performance is obtained as follows: First, using the sample of new hires in their first year at DCPS, fit a regression similar to Table 6 Column 1, except that covariates include each individual component of the academic index and screening scores index separately as in Appendix Table 6 Column 2. The specification also includes subject-taught by year fixed effects. Second, the estimated coefficients from that regression are applied to the applicant sample. This predicted performance measure does not include differences between the subject-taught by year fixed effect groups.

Appendix Table 1—DCPS teacher characteristics

| | Hired before 2011 | | New hires, first year on the job | | | |
|---|---|---|---|---|---|---|
| | First year in data | | Non TeachDC | | TeachDC | |
| | Obs. | Mean (st.dev.) | Obs. | Mean (st.dev.) | Obs. | Mean (st.dev.) |
| Female | 2,920 | 0.76 | 842 | 0.75 | 930 | 0.75 |
| Race/ethnicity | 2,704 | | 380 | | 826 | |
| Black | | 0.60 | | 0.39 | | 0.44 |
| White | | 0.32 | | 0.42 | | 0.47 |
| Hispanic | | 0.04 | | 0.11 | | 0.05 |
| Asian | | 0.04 | | 0.08 | | 0.01 |
| Other | | 0.01 | | 0.00 | | 0.03 |
| Age | 2,914 | 42.32 | 820 | 29.97 | 901 | 31.41 |
| School type | 2,917 | | 839 | | 927 | |
| Education center | | 0.17 | | 0.19 | | 0.18 |
| Elementary school | | 0.46 | | 0.38 | | 0.44 |
| Middle school | | 0.09 | | 0.17 | | 0.15 |
| High school | | 0.25 | | 0.23 | | 0.20 |
| Other | | 0.03 | | 0.03 | | 0.03 |
| Final IMPACT score | 2,920 | 315.04 (45.36) | 842 | 290.50 (48.51) | 933 | 297.86 (46.91) |
| Not in DCPS next year | 2,920 | 0.19 | 842 | 0.21 | 933 | 0.21 |
| Not in same school next year | 2,920 | 0.26 | 842 | 0.29 | 933 | 0.31 |

Note: Sample restricted to DCPS teachers with IMPACT scores. Calculations based on one observation per teacher, the first year they appear in the data.

Appendix Table 2—Factor loadings of IMPACT Component Scores

| IMPACT component score | 2011-12 | | | | 2012-13 | | | | 2013-14 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TLF classroom observation | 0.69 | 0.72 | 0.76 | 0.74 | 0.81 | 0.70 | 0.76 | 0.72 | 0.85 | 0.72 | 0.80 | 0.71 |
| Core professionalism | 0.35 | 0.38 | 0.45 | 0.39 | 0.32 | 0.39 | 0.47 | 0.35 | 0.64 | 0.49 | 0.37 | 0.37 |
| Commitment to school community | 0.70 | 0.63 | 0.73 | 0.75 | 0.75 | 0.75 | 0.74 | 0.75 | 0.82 | 0.79 | 0.78 | 0.73 |
| Teacher assessed student learning | 0.48 | 0.42 | 0.41 | 0.50 | 0.57 | 0.60 | 0.63 | 0.52 | 0.54 | 0.64 | 0.55 | 0.52 |
| Value-added reading | 0.67 | 0.19 | | | 0.47 | 0.17 | | | 0.27 | 0.14 | | |
| Value-added math | 0.67 | | 0.38 | | 0.58 | | 0.44 | | 0.16 | | 0.39 | |

| IMPACT component score | 2014-15 | | | | 2015-16 | | | | 2016-17 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TLF classroom observation | | | | 0.71 | | | | 0.68 | 0.73 | 0.76 | 0.77 | 0.75 |
| Core professionalism | | | | 0.38 | | | | 0.39 | 0.35 | 0.41 | 0.55 | 0.40 |
| Commitment to school community | | | | 0.74 | | | | 0.77 | 0.76 | 0.77 | 0.76 | 0.77 |
| Teacher assessed student learning | | | | 0.50 | | | | 0.52 | 0.15 | 0.57 | 0.55 | 0.55 |
| Value-added reading | | | | | | | | | 0.71 | 0.32 | | |
| Value-added math | | | | | | | | | 0.37 | | 0.29 | |

Appendix Table 3: Pairwise correlations of IMPACT component scores

| | TLF | CP | CSC | TAS | VA read | VA math |
|---|---|---|---|---|---|---|
| TLF classroom observation | 1 | | | | | |
| Core professionalism (CP) | 0.27 | 1 | | | | |
| Commitment to school community (CSC) | 0.63 | 0.34 | 1 | | | |
| Teacher assessed student learning (TAS) | 0.40 | 0.19 | 0.41 | 1 | | |
| Value-added reading | 0.31 | 0.14 | 0.21 | 0.23 | 1 | |
| Value-added math | 0.23 | 0.10 | 0.19 | 0.16 | 0.58 | 1 |

Note: Sample includes all DCPS teachers 2011-12 through 2016-17. Maximum number of teacher-by-year observations for a given correlation is 17,889.

Appendix Table 4—Offers

| | (1) | (2) | (3) |
|---|---|---|---|
| Undergrad GPA (std) | 0.013* | -0.008 | -0.005 |
| | (0.006) | (0.005) | (0.005) |
| SAT/ACT math+verbal (std) | 0.006 | -0.008 | -0.010 |
| | (0.007) | (0.006) | (0.006) |
| Barron's Rank (linear 0-5) | 0.008+ | -0.003 | -0.001 |
| | (0.005) | (0.005) | (0.005) |
| Master's degree or higher | 0.005 | -0.011 | -0.009 |
| | (0.011) | (0.010) | (0.011) |
| Academics index | 0.009 | -0.009+ | |
| | (0.005) | (0.005) | |
| PCK written test (std) | 0.068** | 0.015* | 0.014+ |
| | (0.007) | (0.006) | (0.008) |
| Interview (std) | 0.106** | 0.033** | 0.026* |
| | (0.010) | (0.010) | (0.010) |
| Audition (std) | 0.189** | 0.076** | 0.068** |
| | (0.014) | (0.019) | (0.019) |
| Screening scores index | 0.091** | 0.035** | |
| | (0.008) | (0.008) | |
| Years prior experience | | | |
| 1 to 2 | 0.035* | 0.027+ | 0.027* |
| | (0.015) | (0.014) | (0.014) |
| 3 to 5 | 0.075** | 0.072** | 0.072** |
| | (0.016) | (0.014) | (0.014) |
| 6 to 10 | 0.078** | 0.074** | 0.074** |
| | (0.016) | (0.014) | (0.014) |
| 11 or more | 0.056** | 0.071** | 0.074** |
| | (0.016) | (0.015) | (0.016) |
| Location of undergrad or grad school | | | |
| DC | 0.066** | 0.066** | 0.061** |
| | (0.017) | (0.016) | (0.016) |
| Maryland or Virginia | 0.016 | 0.027* | 0.029** |
| | (0.012) | (0.011) | (0.011) |
| Recommended-pool by year FE | | √ | √ |
| Adjusted R-squared | | | 0.204 |
| F-statistic subject-applied by year FE | | | 1.79 |
| p-value | | | 0.000 |
| F-statistic recommended-pool by year FE | | | 67.8 |
| p-value | | | 0.000 |

Note: Estimates from linear regressions with 5,082 observations from 2012 and 2013, where an indicator for being offered a job is the dependent variable. In columns 1-2 each group of coefficients separated by a solid line are estimates from a separate regression. Columns 3 reports estimates from a single regression. Each specification includes year-by-subject-applied fixed effects. Location indicators are mutually exclusive, applicants with multiple locations coded based on location nearest DC. The recommended-pool by year FE include two mutually exclusive indicators: (i)-(ii) applicants who pass the final audition stage in 2012, and 2013 respectively. The left-out category is all other applicants. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Appendix Table 5—Offers and hiring by subject area applied to teach

| | Offers | | | | Hires | | | |
|---|---|---|---|---|---|---|---|---|
| | Elem. and ECE | MS/HS core | Special education | Other specialties | Elem. and ECE | MS/HS core | Special education | Other specialties |
| | (1) | (2) | (2) | (4) | (5) | (6) | (7) | (8) |
| Academics index | -0.010 | -0.006 | -0.024+ | -0.028* | -0.003 | -0.016* | -0.031* | -0.025* |
| | (0.009) | (0.009) | (0.014) | (0.013) | (0.007) | (0.007) | (0.012) | (0.011) |
| Screening scores index | 0.038** | 0.032* | 0.038+ | 0.035+ | 0.024* | 0.020+ | 0.029 | 0.028+ |
| | (0.014) | (0.015) | (0.020) | (0.019) | (0.010) | (0.011) | (0.020) | (0.015) |
| Years prior experience | | | | | | | | |
| 1 to 2 | 0.028 | 0.002 | 0.081+ | -0.017 | 0.047* | 0.047* | 0.063 | -0.025 |
| | (0.024) | (0.024) | (0.047) | (0.037) | (0.020) | (0.021) | (0.041) | (0.029) |
| 3 to 5 | 0.085** | 0.075** | 0.027 | 0.035 | 0.033+ | 0.064** | 0.019 | -0.018 |
| | (0.026) | (0.023) | (0.037) | (0.037) | (0.019) | (0.019) | (0.033) | (0.028) |
| 6 to 10 | 0.109** | 0.031 | 0.046 | 0.034 | 0.074** | 0.055** | 0.057 | -0.003 |
| | (0.026) | (0.024) | (0.038) | (0.035) | (0.021) | (0.021) | (0.036) | (0.030) |
| 11 or more | 0.123** | 0.029 | 0.049 | 0.029 | 0.038+ | 0.023 | 0.044 | 0.010 |
| | (0.032) | (0.025) | (0.040) | (0.038) | (0.021) | (0.021) | (0.038) | (0.032) |
| Location of undergrad or grad school | | | | | | | | |
| DC | 0.095** | 0.064* | 0.085+ | -0.007 | 0.078** | 0.038+ | 0.045 | 0.057+ |
| | (0.030) | (0.028) | (0.046) | (0.037) | (0.024) | (0.022) | (0.038) | (0.034) |
| Maryland or Virginia | 0.026 | 0.052** | -0.033 | 0.057+ | 0.024 | 0.028+ | 0.006 | 0.038 |
| | (0.020) | (0.020) | (0.029) | (0.029) | (0.015) | (0.017) | (0.028) | (0.023) |
| Rec-pool by year FE | √ | √ | √ | √ | √ | √ | √ | √ |
| Adjusted R-squared | 0.225 | 0.211 | 0.183 | 0.163 | 0.218 | 0.218 | 0.230 | 0.180 |
| F-statistic rec-pool by year FE | 50.0 | 28.5 | 16.8 | 24.3 | 28.7 | 23.2 | 10.6 | 14.5 |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Number of observations | 1667 | 1463 | 657 | 909 | 2441 | 2059 | 710 | 1208 |

Note: For Columns 1-4, the details of estimation are identical to Table 4 Column 2, except that each column above is estimated with a subsample defined in the column headers. For Columns 5-8 estimation follows Table 4 Column 4. "Middle/High school core subjects" includes English, math, sciences, and social studies. "Special education" includes teachers specializing in special education at any grade level. "Other specialties" includes arts, foreign languages, physical education and health, English language learners at any grade level, and others.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Appendix Table 6—Job performance, additional details

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Undergrad GPA (std) | 0.220** | 0.142** | 0.150** | 0.140** | 0.100* |
| | (0.029) | (0.029) | (0.030) | (0.029) | (0.044) |
| SAT math+verbal (std) | 0.166** | 0.034 | 0.033 | -0.010 | 0.073 |
| | (0.032) | (0.032) | (0.032) | (0.030) | (0.046) |
| Barron's Rank (linear 0-5) | 0.141** | 0.102** | 0.103** | 0.088** | 0.026 |
| | (0.023) | (0.022) | (0.022) | (0.020) | (0.028) |
| Master's degree or higher | 0.214** | 0.139** | 0.137** | 0.129** | 0.155* |
| | (0.053) | (0.052) | (0.052) | (0.046) | (0.074) |
| Academics index | 0.250** | | | | |
| | (0.027) | | | | |
| PCK written test (std) | 0.241** | 0.144** | 0.145** | 0.101** | 0.076 |
| | (0.041) | (0.040) | (0.040) | (0.038) | (0.058) |
| Interview (std) | 0.219** | 0.166** | 0.176** | 0.176** | 0.119* |
| | (0.043) | (0.041) | (0.043) | (0.040) | (0.054) |
| Audition (std) | 0.168** | 0.136** | 0.130** | 0.104* | 0.125* |
| | (0.045) | (0.045) | (0.048) | (0.044) | (0.059) |
| Screening scores index | 0.254** | | | | |
| | (0.030) | | | | |
| Years prior experience | | | | | |
| 1 to 2 | 0.035 | 0.065 | 0.061 | 0.027 | 0.061 |
| | (0.075) | (0.070) | (0.070) | (0.064) | (0.097) |
| 3 to 5 | 0.079 | 0.076 | 0.071 | 0.104 | 0.130 |
| | (0.079) | (0.071) | (0.071) | (0.065) | (0.103) |
| 6 to 10 | -0.061 | -0.025 | -0.024 | 0.031 | 0.022 |
| | (0.077) | (0.073) | (0.074) | (0.066) | (0.104) |
| 11 or more | -0.225* | -0.112 | -0.111 | -0.073 | -0.123 |
| | (0.095) | (0.094) | (0.094) | (0.084) | (0.156) |
| Location of undergrad or grad school | | | | | |
| DC | -0.052 | -0.052 | -0.062 | -0.001 | -0.043 |
| | (0.076) | (0.068) | (0.070) | (0.065) | (0.103) |
| Maryland or Virginia | -0.124* | -0.081 | -0.079 | -0.001 | -0.095 |
| | (0.062) | (0.059) | (0.059) | (0.052) | (0.076) |
| Predicted probability of hire | | | | | -1.002 |
| | | | | | (0.851) |
| Predicted probability of hire ^ 2 | | | | | 1.226 |
| | | | | | (0.854) |
| Recommended-pool by year FE | | | √ | √ | |
| School FE | | | | √ | |
| Adjusted R-squared | | 0.196 | 0.198 | 0.356 | 0.223 |
| F-statistic school FE | | | | 12.204 | |
| p-value | | | | 0.000 | |
| F-statistic excluded instruments | | | | | 27.27 |

Note: The details of estimation are identical to Table 6, except that, as shown above, the regressions above include each component of the academic index and screening scores index as a separate covariate.
+ indicates p < 0.10, * 0.05, and ** 0.01

Appendix Table 7—Academics index and school achievement

| | Performance | |
|---|---|---|
| | (1) | (2) |
| Academics index | 0.233** | 0.167** |
| | (0.068) | (0.060) |
| School achievement level | -1.086** | |
| | (0.118) | |
| Academics index * | -0.093 | 0.016 |
| School achievement level | (0.114) | (0.106) |
| Screening scores index | 0.272** | 0.234** |
| | (0.040) | (0.039) |
| Years prior experience | | |
| 1 to 2 | 0.083 | 0.043 |
| | (0.077) | (0.074) |
| 3 to 5 | 0.169* | 0.147+ |
| | (0.086) | (0.077) |
| 6 to 10 | 0.074 | 0.085 |
| | (0.077) | (0.080) |
| 11 or more | -0.106 | -0.101 |
| | (0.105) | (0.105) |
| Location of undergrad or grad school | | |
| DC | 0.000 | 0.041 |
| | (0.080) | (0.081) |
| Maryland or Virginia | -0.031 | -0.008 |
| | (0.067) | (0.064) |
| Recommended-pool by year FE | √ | √ |
| School FE | | √ |
| Adjusted R-squared | 0.230 | 0.352 |
| F-statistic recommended-pool by year FE | 0.6 | 1.1 |
| p-value | 0.746 | 0.371 |

Note: The details of estimation are identical to Table 6, except that the above specifications also include independent variables for school achievement level (omitted in the regressions with school FE) and the interaction of school achievement level and the new hire's academics index. School achievement level is the proportion of students in the hiring school who score proficient or advanced on the 2010-11 DC-CAS.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

Appendix Table 8—Applicants with all three screening scores

| | Hire | | Offer (12-13) | | Performance | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Academics index | -0.032+ | -0.037* | -0.032+ | -0.038* | 0.128** | 0.128** |
| | (0.017) | (0.016) | (0.019) | (0.018) | (0.042) | (0.043) |
| Screening scores index | 0.111** | 0.029+ | 0.136** | 0.029 | 0.163** | 0.177** |
| | (0.014) | (0.016) | (0.016) | (0.019) | (0.038) | (0.042) |
| Years prior experience | | | | | | |
| 1 to 2 | 0.023 | 0.027 | -0.001 | 0.003 | 0.078 | 0.095 |
| | (0.042) | (0.041) | (0.046) | (0.045) | (0.098) | (0.100) |
| 3 to 5 | 0.164** | 0.151** | 0.151** | 0.141** | 0.156+ | 0.166+ |
| | (0.046) | (0.043) | (0.053) | (0.048) | (0.093) | (0.093) |
| 6 to 10 | 0.143** | 0.141** | 0.141** | 0.141** | 0.014 | 0.038 |
| | (0.047) | (0.045) | (0.050) | (0.048) | (0.098) | (0.100) |
| 11 or more | 0.067 | 0.082 | 0.016 | 0.037 | -0.220 | -0.213 |
| | (0.058) | (0.052) | (0.062) | (0.057) | (0.172) | (0.175) |
| Location of undergrad or grad school | | | | | | |
| DC | 0.042 | 0.042 | 0.065 | 0.060 | -0.027 | -0.019 |
| | (0.045) | (0.043) | (0.049) | (0.046) | (0.101) | (0.102) |
| Maryland or Virginia | 0.046 | 0.044 | 0.080* | 0.072+ | -0.075 | -0.053 |
| | (0.036) | (0.035) | (0.040) | (0.038) | (0.079) | (0.080) |
| Recommended-pool by year FE | | √ | | √ | | √ |
| Adjusted R-squared | 0.079 | 0.162 | 0.104 | 0.209 | 0.235 | 0.238 |
| F-statistic subject-applied by year FE | 1.03 | 0.93 | 0.73 | 0.75 | | |
| p-value | 0.406 | 0.641 | 0.908 | 0.888 | | |
| F-statistic rec-pool by year FE | | 50.3 | | 69.9 | | 4.380 |
| p-value | | 0.000 | | 0.000 | | 0.001 |

Note: For Columns 1-2 above, the details of estimation are identical to Table 3 Columns 4-5, except that the sample has been limited to applicants with all three screening scores: PCK, interview, and audition. Similarly, Columns 3-4 above follow Table 4 Columns 3-4, and Columns 5-6 above follow Table 6 Columns 1-2.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

## Appendix B: Literature on the Teacher Hiring Process

Outside of economics, there is a significant literature on teacher hiring. These studies focus mostly on either (a) the characteristics that principals look for in job applicants and (b) the processes used by schools to recruit, screen, and select teachers. Here we give a general sense of these literatures and their findings; for more detailed reviews see Pounder (1989), Rutledge et al. (2008), or Mason and Schroeder (2010).

Studies of the values principals place on teacher applicant characteristics are based on qualitative interviews and surveys, typically with small samples drawn from either one district or a limited geographic area (e.g., Place and Kowalski 1993, Abernathy et al. 2001, Harris et al. 2007, Cannata and Engel 2012). These analyses generally indicate that principals place greater weight on personal traits (e.g., "honesty", "good character", "ability to work with peers", "respect" or "compassion" for students) that may be more difficult to assess than credentials like academic achievement or years of prior teaching experience.

No nationally representative study on methods used for teacher hiring exists. However, a number of studies, spanning many years and various geographic areas, provide a fairly consistent picture. The two methods employed in these studies are either to ask school district administrators about their hiring practices or to survey teachers about their experiences being hired for their most recent job (Liu and Kardos 2002, Liu and Moore-Johnson 2007).

Applicants almost always submit written applications with information including a resume and proof of certification, as well as transcripts and recommendation letters. From there, a subset of applicants is invited for in-person interview. These surveys also indicate that teachers are usually interviewed more than once as they progress toward being hired.

Submission of writing samples, a portfolio of work, or delivering a sample lesson are all far less common than the in-person interview. None of the early studies we reviewed mentioned any written evaluations other than a cover letter and they all report that a small fraction (typically less than 15 percent) of districts observed applicants teaching a lesson prior to hire.[1] More recently, Strauss (1998) reports that roughly 25% of districts surveyed in Pennsylvania solicit writing samples and roughly one third request teacher exam scores (e.g., National Teacher Examination or Praxis Series). Balter and Duncombe (2005), surveying New York State school districts, report that 60% require a writing sample, 30% require a teaching portfolio, and two thirds require certification exam scores.[2] These surveys also report between 40 and 50 percent of districts using a sample classroom presentation, which may indicate a trend toward greater use. However, surveys of teachers (Liu and Kardos 2002, Liu and Moore-Johnson 2007) across several states typically find only about 15% giving a sample lesson prior to being hired.[3] One caveat to this conclusion is that student teachers and teachers' aides who are hired to teach full-time, will likely have been observed while teaching, albeit outside of the formal hiring process.[4]

A large literature in applied (or "industrial") psychology examines the power of interviews to extract reliable and accurate information about the future success of potential employees. Much of the early research in this field exposed low reliability and validity (see Schmitt 1976), but more recent work demonstrates that structured interviews (i.e., pre-selected

[1] See Neely (1957), Diekrager (1969), Nalley (1971), Hansen (1976), and Luthy (1982).
[2] In addition to letters of recommendation, Pennsylvania districts reported placing high weight on college major and grade point average (but low weight on test scores, essays, or institution attended) when deciding whom to interview. In New York, recommendations and college major are also given high weight in screening prior to the interview, but low weights are given to grade point average, institution attended, and scores on certification exams (or other screening tests).
[3] The fraction of teachers who taught a sample lesson was 6.5 percent in California, 14 percent in Florida, 14.6 percent in Michigan, 19.6 percent in Massachusetts, and 23 percent in New Jersey.
[4] Liu and Moore-Johnson (2007) report that 20 percent of teachers worked in their current schools in some capacity before they were hired, and Strauss (1998) finds that about one third of school districts try to fill full-time teaching positions with current substitutes or part-time teachers.

questions with rubrics for coding answers) can predict outcome variables such as evaluations of employee performance by supervisors (see Arvey and Campion 1982, Hunter and Hunter 1984, Motowidlo et al. 1990, McDaniel et al. 1994).

A small set of studies focus on interviews for teachers specifically. However, this research typically examines how teacher characteristics (e.g., gender, age) and interview structure (e.g., a single interviewer vs. a panel) affect hiring decisions, and many of these studies use actors instead of actual teachers (e.g., Bolton 1969, Young and Pounder 1986, Young 2005). Few studies test whether interview decisions predict future success in teaching, but there is some evidence, albeit in small samples, of a positive relationship between teachers' interview ratings and supervisor ratings of job performance (Mickler and Solomon 1986) and student achievement gains (Webster 1988).

We know of no study that focuses on the predictive validity of sample lessons done as part of a hiring process.[5] However, the power of job simulations to predict productivity is a well-researched issue in the field of industrial psychology (see Wernimont and Campbell 1968, Hunter and Hunter 1984), and there is a large literature on the relationship between student achievement and observed teacher behaviors—typically measured with trained observers using low-inference coding systems or "rubrics." This research has consistently found positive relationships between observed teacher behavior and student learning outcomes.[6] Nevertheless, while research indicates that effective teachers can be identified through observation, this

[5] Some indirect evidence is presented by Wede (1996), who analyzed data on subjective performance evaluations of teachers from a school district that incorporated a sample lesson as part of its hiring process. Several years later, average evaluations of those hired during this period were not statistically different than those hired in prior years.
[6] Recent work includes Holtzapple (2003), Schacter and Thum (2004), Milanowski (2004), Kimball et al. (2004), Gallagher (2004), Kane et al. (2011), and Kane et al. (2012). Brophy and Good (1984) review earlier research.

evidence comes from full-time teachers in normal classroom settings and may not accurately reflect the evaluation of sample lessons presented during the hiring process.

## References


Abernathy, Tammy V., Al Forsyth and Judith Mitchell. 2001. "The Bridge from Student to Teacher: What Principals, Teacher Education Faculty, and Students Value in a Teaching Applicant." *Teacher Education Quarterly* 28(4): 109-119.

Arvey, Richard D., and James E. Campion. 1982. "The Employment Interview: A Summary and Review of Recent Research." *Personnel Psychology* 35(2): 281-322.

Balter, Dana and William D. Duncombe. 2005. "Teacher Hiring Practices in New York State Districts," Report prepared for the Education Finance Research Consortium.

Bolton, D.L. 1969. "The Effect of Various Information Formats on Teacher Selection Decisions," *American Educational Research Journal*, 6(3): 329-347

Brophy, Jere, and Thomas L. Good. 1986. Teacher Behavior and Student Achievement. In M. C. Wittrock (Ed.), *Handbook of Research on Teaching*. 3rd ed., 238-375, New York: Simon and Schuster.

Cannata, Marisa, and Mimi Engel. 2012. "Does Charter Status Determine Preferences? Comparing the Hiring Preferences of Charter and Traditional Public School Principals." *Education* 7(4): 455-488.

Diekrager, Wayne A. 1969. "Teacher Selection: A Synthesis and Integration of Research Findings," Doctoral Dissertation, University of South Dakota.

Gallagher, H. Alix. 2004. "Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement?" *Peabody Journal of Education* 79(4): 79-107.

Hansen, Cecil Ray. 1976. "Practices and Procedures Used by Selected Utah Public School Districts in the Recruitment and Selection of Teachers." Doctoral Dissertation, Brigham Young University.

Harris, Douglas N., and Tim R. Sass. (2011) "Teacher Training, Teacher Quality and Student Achievement." *Journal of Public Economics* 95(7–8):798-812.

Holtzapple, Elizabeth. 2003. "Criterion-Related Validity Evidence for a Standards-Based Teacher Evaluation System." *Journal of Personnel Evaluation in Education* 17(3): 207-219.

Hunter, John E., and Ronda F. Hunter. 1984 "Validity and Utility of Alternative Predictors of Job Performance." *Psychological Bulletin* 96(1): 72-98.

Kane, Thomas J., Douglas O. Staiger, and Dan McCaffrey. 2012. *Gathering Feedback for Teaching.* Seattle: Bill and Melinda Gates Foundation. Accessed May 2015 at http://www.metproject.org/reports.php.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

Kimball, Steven M., Brad White, Anthony T. Milanowski, and Geoffrey Borman. 2004. "Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County." *Peabody Journal of Education* 79(4): 54-78.

Liu, Edward, and Susan M. Kardos. 2002. *Hiring and Professional Culture in New Jersey Schools.* Cambridge, MA: Project on the Next Generation of Teachers at the Harvard Graduate School of Education.

Liu, Edward, and Susan Moore-Johnson. 2006. "New Teachers' Experiences of Hiring: Late, Rushed, and Information Poor." *Educational Administration Quarterly* 42(3): 324-360.

Mason, Richard W., and Mark P. Schroeder. 2010. "Principal Hiring Practices: Toward a Reduction of Uncertainty." *The Clearing House* 83(5): 186–193.

McDaniel, Michael A., Deborah L. Whetzel, Frank L. Schmidt, and Steven D. Maurer. 1994. "The validity of employment interviews: A comprehensive review and meta-analysis." *Journal of Applied Psychology* 79(4): 599-616.

Mickler, Mary Louise, and Gloria L. Solomon. 1986. "Beyond Credentials in Teacher Selection: A Second Look at the Omaha Teacher Interview." *North Central Association Quarterly* 60(3): 398-402.

Milanowski, Anthony. 2004. "The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence from Cincinnati." *Peabody Journal of Education* 79(4): 33-53.

Motowidlo, Stephen J., Marvin D. Dunnette, and Gary W. Carter. 1990. "An Alternative Selection Procedure: The Low-Fidelity Simulation." *Journal of Applied Psychology* 75(6): 640-647.

Nalley, B. J. 1971. "A Descriptive Survey of the Recruitment and Selection Process of Teachers for the District of Columbia and a Comparison of Procedures Used in Selected School Systems of Comparable Size." Doctoral dissertation, George Washington University, Dissertation Abstracts International, 32, 3626A.

Neely, Melvin E. 1957. "A Survey of Present Procedures in the Selection of Teacher Personnel." Doctoral Dissertation, University of Kansas.

Place, Andrew W., and Theodore J. Kowalski. 1993. "Principal Ratings of Criteria Associated with Teacher Selection," *Journal of Personnel Evaluation in Education* 7: 291-300.

Pounder, Diana G. 1989. "Improving the Predictive Validity of Teacher Selection Decisions: Lessons from Teacher Appraisal." *Journal of Personnel Evaluation in Education* 2(2): 141-150.

Rutledge, Stacey A., Douglas N. Harris, Cynthia T. Thompson, and W. Kyle Ingle. 2008. "Certify, Blink, Hire: An Examination of the Process and Tools of Teacher Screening and Selection." *Leadership and Policy in Schools* 7(3): 237–263.

Schacter, John, and Yeow M. Thum. 2004. “Paying for High- and Low- Quality Teaching.” *Economics of Education Review* 23(4): 411-440.

Schmitt, Neal. 1976. “Social and Situational Determinants of Interview Decisions: Implications for the Employment Interview.” *Personnel Psychology* 29(1): 79-101.

Strauss, R. 1998. “Teacher Preparation and Selection in Pennsylvania,” Research Report to the Pennsylvania State Board of Education.

Webster, William J. 1988. “Selecting Effective Teachers.” *The Journal of Educational Research* 81(4): 245-253.

Wede, Richard J. 1996. “Teacher Selection: Use of Demonstration Lessons” Doctoral Dissertation, Drake University.

Wernimont, Paul F., and John P. Campbell. 1968. “Signs, Samples, and Criteria.” *Journal of Applied Psychology* 52(5): 372.

Young, I.P., Pounder, D.G. 1986. “Salient Factors Affecting Decision Making in Simulated Teacher Selection Interviews,” *Journal of Educational Equity and Leadership*, 5(3):216-233.

Young, I. Phillip. 2005. “Effects of ‘Like Type’ Sex Pairings Between Applicants–Principals and Type of Focal Position Considered at the Screening Stage of the Selection Process.” *Journal of Personnel Evaluation in Education* 18(3): 185-199.

## Appendix C: TeachDC Application Process

We focus on the TeachDC selection process as it occurred from 2011-2013. Each year, from roughly February through July, candidates submit applications to TeachDC. The online application system first collects background information such as applicants' education history, employment experience, and eligibility for licensure. Applicants who don't already hold a DC license and whose credentials make them ineligible to obtain one prior to the start of the school year are not allowed to proceed further, and we do not analyze these ineligible applications.[1]

Following collection of this preliminary information, district officials review applications in several stages. In 2011, there were four stages of evaluation; two written evaluations (general essays and subject-specific assessments), an interview, and a teaching audition. In 2012 and 2013, the general essay was dropped, and applicants were assessed on the remaining three stages. Also, in 2011 many of the teaching auditions were done live in DCPS classrooms, but, due to logistical difficulties, audition videos were used for the remaining 2011 cases and in all of 2012 and 2013.

At the end of each stage, applicants who pass a specified performance threshold are allowed to proceed. Applicants who pass all stages (and a background check) are included in the recommended pool seen online by principals. On average, for those who made it through the process, it took roughly six weeks from the initial application to the pass/fail determination at the final stage.

Appendix Table C1 shows the number of applicants evaluated in each recruiting year and each stage, as well as whether or not they passed the stage and the fraction of applicants hired in

[1] To be licensed in DC, teachers must have a bachelor's degree, complete a teacher preparation program (traditional or alternative), and pass both the PRAXIS I and relevant PRAXIS II exams (or substitute exams). Teachers licensed in another state are also generally eligible for a DC license.

each possible stage outcome. There were roughly 2,500 applicants per year, of which roughly 13 percent were hired into DCPS. Roughly 60-70% of applicants completed the subject-specific written assessment and 30-40% of applicants completed the interview. The number of applicants completing the audition rose significantly after 2011, when teachers were encouraged to submit a video audition.

As mentioned above, applicants did not have to make it into the TeachDC recommended pool in order to be hired into DCPS.[2] However, Table C1 shows that applicants who passed the stages were more likely to be hired than those who did not. Among those applicants who passed the final audition stage and made it into the recommended pool, the fraction hired was 48 percent, 42 percent, and 52 percent in years 2011, 2012, and 2013, respectively.

To give a better sense of how the TeachDC process worked in practice, we briefly summarize the key aspects of each stage during the three years on which we focus. In 2011, applicants first submitted online essays of 200-400 words which were scored by one of several district office reviewers for content and writing quality.[3]

Applicants in all three years took a subject-specific written assessment to assess their pedagogical content knowledge (PCK) and knowledge of instructional practices. Applicants selected a subject area (e.g., art, math, Biology) and level (i.e., elementary, middle, or high school) to which they were applying, and then were asked to complete a subject- and level-

[2] Appendix Table C1 shows that 31% of the 80 candidates who failed the audition stage in 2011 were nonetheless hired by the district. The analogous figures in 2012 and 2013 are notably lower. Based on conversations with DCPS officials, we believe that some of these candidates were actually moved ahead to the recommended pool to increase the choices available to principals. In analyses we do not present (but which are available upon request), we confirm that all of our results are robust to excluding these 80 applicants or to treating them as recommended.

[3] One essay was on instructional strategies for low-performing students, and the other on the use of student achievement data. These essays were scored by on a 4 point scale (in 0.1 point increments), and a composite score was calculated using weights of 40% for the content of each essay and 20% for overall writing quality. As a general rule, applicants proceeded if they achieved a composite score of 2.0 or higher. In addition, DCPS officials selected a random 20% subset of applicants with scores below 2.0 to pass, although applicants with the minimum possible score (1.0 on both essays) were not eligible to be selected.

specific task. Most applicants were asked to read a case-study in which a student demonstrates misunderstanding of the subject matter and to write a 300-400 word essay explaining the nature of the student's misconceptions and describing instructional strategies for resolving them. In 2011 and 2012, applicants for math teaching positions were required to complete the Mathematical Knowledge for Teaching (MKT) test, a multiple choice test intended to measure understanding and skills distinctly valuable to teaching math (Hill et al. 2004).[4] Essay content and writing quality were scored by DCPS personnel and these scores (plus the KMT test score, when applicable) were averaged to determine whether the applicant passed to the next stage.

Applicants who passed the subject-specific essay stage were invited for a 30-minute interview and to submit a 10-minute demonstration lesson. Interviews were conducted by the same DCPS personnel who scored the subject-specific essays, as well as several "Teacher Selection Ambassadors" (TSAs). TSAs were highly rated DCPS teachers who received training from DCPS staff in order to assist with the TeachDC selection process.[5]

The demonstration or "mini" lesson could be done in person or submitted by video. Applicants were allowed to choose the topic and had the option to provide lesson materials. DCPS officials scored applicant performance according to selected dimensions of the Teaching and Learning Framework (TLF), the same rubric used to measure classroom performance under

[4] Elementary school applicants wrote an essay assessing content knowledge in English language arts in addition to taking the MKT test. Applicants for middle school math positions in these two years completed the MKT but did not have to complete an additional essay. In 2013, DCPS did not administer the MKT assessment, instead relying on essays alone to evaluate each candidate's content knowledge.

[5] Interviews could be done in person or over the phone, and applicants were asked to respond to a series of structured questions covering five areas: track record of success, response to challenges, contribution to work environment, ownership of high expectations, and continuous learning. For example, under "response to challenges," interviewees were asked, "tell me about the most significant behavior challenge that you've encountered with a student (or group)," with follow-up questions like "what did you do first to address the challenge," "what was the result," and "what ultimately happened." Applicants' responses were scored on a 4-point scale using a detailed rubric.

the DCPS IMPACT teacher evaluation system, which we describe in more detail below.[6] Applicant performance on the mini-lesson and interview were combined to yield a final score, and applicants scoring above a specified threshold, which varied somewhat across years, were invited to proceed to the final stage. In 2013, DCPS did not require the mini-lesson and applicants were evaluated on the basis of the interview alone.

The final stage in the TeachDC process consisted of a teaching audition in which the applicant taught a complete lesson of approximately 30 minutes. All auditions in 2011 were conducted in DCPS classrooms but were videotaped for evaluation. In 2012, applicants were permitted to submit a videotaped teaching lesson in lieu of the "live" audition, while in 2013 auditions were based completely on video submissions. In each year, DCPS staff and TSAs evaluated the auditions using the same DCPS classroom observation protocol (i.e., the TLF rubric mentioned above), with each audition rated by one TSA.[7]

In addition to the measures used for selection, applicants in 2011 were asked additional questions during the first stage meant to assess the candidate's personality and a commercial teacher selection product (Haberman Star Teacher Pre-Screener). Here we discuss these measures in detail and show estimates of how these predict teacher hiring and performance. These measures were not used in the selection process, but, importantly, applicants were not told

[6] Applicants receive a score of 1-4 in five areas: lead well-organized objective-driven lessons, explain content clearly, engage students in learning at all levels, check for student understanding, and maximize instructional time. The scoring rubric is quite detailed and the current version can be found at: http://dcps.dc.gov/page/impact-overview. To provide an example of how scores are anchored, some of the language describing a "4" in "maximize instructional time" includes "routines, procedures, and transitions are orderly, efficient, and systematic with minimal prompting from the teacher." By contrast, a score of "1" is described by "routines or procedures are not evident or generally ineffective; the teacher heavily directs activities and transitions."

[7] Applicants received a score from 1-4 in the same five areas described in the previous footnote. In 2013, approximately 15% of interviews and 30% of the auditions were checked by a DCPS staff member as part of a "random audit" to assess the reliability of TSA ratings. The correlation between the average scores initially assigned and those after review was 0.87 for interviews, although 45% had at least one component score changed and 17% had the final recommendation overturned. Only 20% of reviewed auditions had any component score changed, leading to roughly 10% of reviewed auditions having the final recommendation overturned.

explicitly that these items were "low-stakes", so these data are likely indicative of responses that DCPS would receive if they were to be used in the selection process.[8]

Applicants answered 50 multiple-choice questions from the Haberman Star Teacher Pre-Screener (Haberman, 1993), a commercial teacher applicant screening instrument. Used by a number of large urban school districts throughout the U.S., the Haberman Pre-Screener is intended to provide school officials with guidance on how effective a particular candidate is likely to be in an urban classroom. Prior research has indicated a positive relationship between Haberman scores and teacher performance in the classroom (Rockoff et al. 2011).[9]

Applicants also answered multiple-choice questions to measure the "Big Five" personality traits (Costa and McCrae, 1992) and Grit, defined as "the tendency to sustain interest in and effort toward very long-term goals" (Duckworth and Quinn, 2009).[10] While our intention was to examine measures of the Big Five and Grit, a factor analysis (see Appendix Table C2) reveals that applicants' answers are inconsistent with the instruments' designs. The only trait from these surveys that aligns well with a cohesive set of personality questions is Extroversion. All questions other than Extroversion line up along two factors corresponding to whether the

[8] Prior to this entire section, applicants were informed that some of the questions were part of a pilot program, but were not told which items were part of the pilot and which were not.

[9] This assessment was developed by interviewing teachers thought to be highly effective and designing questions to capture their attitudes and beliefs. The Haberman Foundation also produces an interview protocol and scoring rubric which is intended to assist district officials in identifying individuals likely to be effective urban school teachers, although this protocol was not used in DCPS during the period of our study. The average score (out of 50) for 2011 TeachDC applicants was 34.2, with a standard deviation of 4.7, and similar to the average score of 31.9 (standard deviation 4.8) found by Rockoff et al. (2011) for a sample of recently hired NYC math teachers.

[10] Personality traits were measured using a shortened version of the Big Five Inventory (John, Donahue, and Kentle 1991) in which applicants express their degree of agreement with how a phrase (e.g., "I am talkative") describes them. The 16 items focused mostly on Extroversion (5 questions) and Conscientiousness (5 questions), two traits linked to job performance in earlier studies (Barrick and Mount, 1991; Rockoff et al., 2011), with less emphasis on Agreeableness (2 questions), Neuroticism (2 questions), or Openness to New Experience (2 questions). Grit was measured using a similar instrument developed by Duckworth and Quinn (2009) with eight items, such as "is not discouraged by setbacks" and "has difficulty maintaining focus on projects that take more than a few months." The definition of Grit is provided at: https://sites.sas.upenn.edu/duckworth , accessed on March 17, 2014.

question was normally scored (e.g., measuring conscientiousness, the item "Is a reliable worker") or reverse scored (e.g., measuring conscientiousness, the item "Tends to be disorganized"). We believe this was because the questions were asked on a job application rather than a survey, and candidates likely "faked" their responses to appear more attractive.[11] Hence, in the analysis below, we include three personality measures (Extroversion, "Positive Spin", and "Negative Spin"), and the Haberman test score in regressions of DCPS hiring and teacher performance.

Appendix Table C3 shows pairwise correlations for the additional measures collected for the 2011 application cohort. In general, these measures are not at all highly correlated with the other application performance measures. There is a modest correlation between extraversion and interview and audition scores (0.14 and 0.13, respectively) and the Haberman score has small positive correlations of roughly 0.2 with the academic measures and the PCK assessment.

Extraversion and the Haberman Index are both positively associated with the likelihood of being hired (see Appendix Table C4), although the Haberman coefficient becomes small and insignificant once we condition on being in the TeachDC recommended pool. When interpreting these results it is important to remember that these measures were not made available to principals, because DCPS officials were uncertain about their usefulness, but this fact was not told to applicants in order for the data collection to reflect normal conditions.

Turning to the results for job performance, the most interesting result to emerge is that coefficient on the Haberman score is large, positive, and significantly associated with teacher performance (Appendix Table C5). Specifically, a one standard deviation increase in an

[11] A comparison with responses of roughly 400 recently hired New York City math teachers on a low-stakes survey of the Big Five (Rockoff et al. 2011, Table 2) supports this notion. NYC teachers reported levels (on a 5 point scale) of 4.11, 4.04, and 3.85 for, respectively, Agreeableness, Conscientiousness, and Openness to New Experiences. Each had a standard deviation of about 0.5. In stark contrast, the 2011 TeachDC applicants' average reported Agreeableness, Conscientiousness, and Openness to New Experiences were 4.63, 4.67, and 4.66. For other evidence on self-report bias in this context, see Mueller-Hanson et al. (2003) and Donovan et al. (2014).

applicant's score on the Haberman Index is associated with a 0.27 standard deviation increase in measured effectiveness, even after controlling for reaching the recommended pool.

## References


Barrick, Murray R., and Michael K. Mount. 1991. "The Big Five Personality Dimensions and Job Performance: A Meta-Analysis." *Personnel Psychology* 44(1), 1-26.

Costa, Paul T., and Robert R. McCrae. 1992. *Neo PI-R Professional Manual*. Lutz, FL: Psychological Assessment Resources.

Donovan, John J., Stephen A. Dwight, Dan Schneider. 2014. "The Impact of Applicant Faking on Selection Measures, Hiring Decisions, and Employee Performance." *Journal of Business Psychology* 29: 479–493.

Duckworth, Angela Lee, and Patrick D. Quinn. "Development and Validation of the Short Grit Scale (GRIT–S)." *Journal of Personality Assessment* 91.2(2009): 166-174.

Haberman, Martin. 1993. "Predicting the Success of Urban Teachers (The Milwaukee Trials)." *Action in Teacher Education* 15(3): 1-5.

Hill, Heather. C., Stephen G. Schilling, and Deborah L. Ball. 2004. "Developing Measures of Teachers' Mathematics Knowledge for Teaching." *Elementary School Journal* 105: 11–30.

John, Oliver P., Eileen M. Donahue, and Robert L. Kentle. 1991. *The "Big Five" Inventory—Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Mueller-Hanson, Rose, Eric D. Heggestad, and George C. Thornton III. 2003. "Faking and Selection: Considering the Use of Personality from Select-In and Select-Out Perspectives." *Journal of Applied Psychology* 88(2): 348–355

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2011. "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy* 6(1): 43-74.

Appendix Table C1—Applicant progress through TeachDC process

| | | 2011 applicants | | 2012 applicants | | 2013 applicants | |
|---|---|---|---|---|---|---|---|
| | | # | Fraction hired | # | Fraction hired | # | Fraction hired |
| Eligible but Initial Stage Incomplete: | | 174 | 0.13 | 787 | 0.10 | 1,041 | 0.02 |
| General Essay | Failed this Stage: | 228 | 0.04 | | | | |
| | Incomplete Next Stage: | 362 | 0.09 | | | | |
| Content Knowledge | Failed this Stage: | 530 | 0.06 | 622 | 0.05 | 260 | 0.02 |
| | Incomplete Next Stage: | 314 | 0.09 | 304 | 0.09 | 150 | 0.06 |
| Interview + Sample Lesson | Failed this Stage: | 239 | 0.09 | 283 | 0.05 | 184 | 0.03 |
| | Incomplete Next Stage: | 269 | 0.32 | 39 | 0.18 | 302 | 0.09 |
| Teaching Audition | Failed this Stage: | 80 | 0.31 | 100 | 0.01 | 156 | 0.04 |
| | Passed Stage: | 164 | 0.48 | 392 | 0.42 | 462 | 0.52 |

Note: Authors' calculations. "Stage reached" is the highest stage in which the data include a score or pass/fail determination.

Appendix Table C2—Factor analysis of personality and grit questions

| | Positive Spin | Negative Spin | Extraversion |
|---|---|---|---|
| + BFI Q3 | **0.43** | -0.10 | 0.09 |
| + BFI Q8 | **0.44** | -0.04 | -0.02 |
| + BFI Q15 | **-0.41** | -0.13 | 0.06 |
| + BFI Q17 | **0.55** | 0.01 | -0.06 |
| + BFI Q23 | **0.39** | 0.22 | 0.00 |
| + BFI Q26 | **0.45** | 0.00 | -0.01 |
| + BFI Q28 | **0.48** | -0.02 | 0.03 |
| + Grit Q2 | **0.26** | 0.17 | 0.00 |
| + Grit Q4 | **0.55** | -0.06 | -0.05 |
| + Grit Q7 | **0.41** | 0.28 | -0.11 |
| + Grit Q8 | **0.60** | 0.11 | -0.11 |
| - BFI Q5 | 0.09 | **0.43** | 0.01 |
| - BFI Q11 | 0.09 | **0.46** | -0.03 |
| - BFI Q12 | 0.20 | **0.17** | 0.11 |
| - BFI Q22 | 0.00 | **0.69** | 0.01 |
| - BFI Q24 | 0.17 | **0.20** | -0.07 |
| - BFI Q25 | 0.20 | **0.26** | 0.02 |
| - BFI Q29 | -0.07 | **-0.42** | -0.17 |
| - Grit Q1 | -0.04 | **0.64** | 0.01 |
| - Grit Q3 | -0.04 | **0.62** | 0.08 |
| - Grit Q5 | -0.01 | **0.51** | 0.04 |
| - Grit Q6 | 0.08 | **0.54** | -0.06 |
| Extraversion Q1 | 0.04 | -0.18 | **0.60** |
| Extraversion Q4 | -0.15 | 0.04 | **0.60** |
| Extraversion Q7 | **0.40** | -0.05 | **0.34** |
| Extraversion Q10 | **0.43** | -0.06 | **0.34** |
| Extraversion Q13 | -0.11 | 0.07 | **0.71** |
| Extraversion Q16 | 0.19 | -0.01 | **0.34** |
| Extraversion Q19 | -0.16 | 0.25 | **0.62** |
| Extraversion Q21 | 0.26 | -0.08 | **0.59** |

Note: This table presents the results of a factor analysis on items used to assess the Big Five personality traits as well as Duckworth's Grit measure, where factors having an eigenvalue greater than 1 were retained. Factor weights are given with a Promax Rotation. Items with "+" are positively scored, those with "-" are negatively scored. BFI refers to big five items that do not pertain to extraversion. Note that items BFI 15 and 29 relate to neuroticism, and hence are inversely related to the underlying factors.

Appendix Table C3—Pairwise correlations of applicant characteristics and scores

| | SAT | GPA | Barr. | Exper. | PCK | Interv. | Aud. | Personality questions: Extrov. | Personality questions: Pos. | Personality questions: Neg. | Haber-man |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Personality questions | | | | | | | | | | | |
| Extroversion | 0.07 | 0.02 | 0.06 | -0.12 | 0.06 | 0.14 | 0.13 | 1 | | | |
| Positive spin | -0.04 | 0.01 | -0.01 | 0.04 | -0.02 | 0.04 | -0.01 | 0.28 | 1 | | |
| Negative spin | -0.05 | 0.01 | -0.04 | 0.04 | -0.02 | 0.04 | -0.04 | 0.26 | 0.70 | 1 | |
| Haberman total score | 0.21 | 0.20 | 0.19 | -0.14 | 0.20 | 0.12 | 0.01 | 0.13 | 0.11 | 0.07 | 1 |

Note: This table is a companion to Table 2. Pairwise correlations of applicant characteristics and scores. Maximum observations for a cell is 2,360.

Appendix Table C4—Hiring
Additional characteristics from 2011 applicants

| | Characteristics separately | | Characteristics simultaneously | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Positive spin factor (std) | -0.015 | -0.013 | -0.016 | -0.013 |
| | (0.012) | (0.011) | (0.012) | (0.011) |
| Negative spin factor (std) | 0.006 | 0.011 | 0.006 | 0.010 |
| | (0.011) | (0.011) | (0.011) | (0.011) |
| Big Five Index: Extroversion (std) | 0.030** | 0.019** | 0.028** | 0.018* |
| | (0.008) | (0.007) | (0.008) | (0.007) |
| Haberman total score (std) | 0.018* | 0.005 | 0.012 | 0.002 |
| | (0.007) | (0.007) | (0.008) | (0.007) |
| General teaching essay (std) | 0.018* | 0.001 | 0.015+ | 0.001 |
| | (0.008) | (0.008) | (0.009) | (0.008) |
| Recommended-pool FE | | √ | | √ |
| Number of observations | | | 2,360 | 2,360 |
| Adjusted R-squared | | | 0.027 | 0.133 |
| *F*-statistic subject-applied FE | | | 1.95 | 1.12 |
| p-value | | | 0.002 | 0.300 |
| *F*-statistic recommended-pool FE | | | | 143 |
| p-value | | | | 0.000 |

Note: Estimates from an LPM with 2,360 observations (all from 2011) where being hired is the dependent variable. In Columns 1-2 each group of coefficients separated by a solid line are estimates from a separate regression. Columns 3-4 each report estimates from a single regression. Each specification includes year-by-subject-applied fixed effects. The recommended-pool fixed effects include two mutually exclusive indicators: (i) applicants who pass the audition in 2011, and (ii) applicants who pass the interview but not the audition in 2011 (see the text for more details). The left out category is all other applicants. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate.
+ indicates p < 0.10, * 0.05, and ** 0.01

Appendix Table C5—Job performance
Additional characteristics from 2011 applicants

| | (1) | (2) |
|---|---|---|
| Positive spin factor (std) | 0.046 | 0.022 |
| | (0.055) | (0.055) |
| Negative spin factor (std) | -0.048 | -0.018 |
| | (0.067) | (0.067) |
| Big Five Index: Extroversion (std) | 0.005 | -0.006 |
| | (0.053) | (0.054) |
| Haberman total score (std) | 0.268** | 0.248** |
| | (0.051) | (0.051) |
| General teaching essay (std) | 0.208** | 0.174** |
| | (0.065) | (0.065) |
| Recommended-pool FE | | √ |

Note: Estimates from least squares regressions with 744 teacher-by-year observations, and 314 unique teachers (hired in 2011 only). The dependent variable is job performance measured by the first predicted factor from a factor analysis of IMPACT evaluation component scores, standardized. Each group of coefficients separated by a solid line are estimates from a separate regression. Each specification includes year-by-subject-taught fixed effects, and indicators for second year in the district and third year in the district. The recommended-pool fixed effects include two mutually exclusive indicators: (i) applicants who pass the audition in 2011, and (ii) applicants who pass the interview but not the audition in 2011 (see the text for more details). The left out category is all other applicants. When a covariate is missing for an observation, we set the value to zero and include an indicator = 1 for anyone missing that covariate. Clustered (teacher) standard errors in parentheses.
+ indicates $p < 0.10$, * 0.05, and ** 0.01

## Appendix D: DCPS Performance Evaluation (DCPS IMPACT)

Each DCPS teacher's performance for the previous school year is summarized in a single "IMPACT" score which determines personnel decisions ranging from termination to significant salary increases. An IMPACT score is composed of several measures which vary depending on the grade(s) and subject(s) to which the teacher is assigned. Our data include final IMPACT scores and all components for all district teachers in the school years 2011-12 through 2015-16.

The first component of IMPACT is based on measures of student learning. For math and reading teachers in grades 4 to 10 (4 to 8 in 2011-12), this includes an individual value-added score (IVA) based on the DC Comprehensive Assessment System (DC-CAS) standardized tests. These teachers, known as "Group 1", represent about 15 percent of DCPS teachers. All teachers are evaluated with a "Teacher Assessed Student Achievement" score (TAS), which are teacher-specific learning goals, designed and assessed by the teacher herself. [1] Additionally, in 2011-12, 5 percent of teachers' final IMPACT score is a measure of school value-added on DC-CAS tests.

The second component is a classroom observation score. Each teacher is typically observed five times annually, three times by the principal and twice by a "master educator" (i.e., an experienced teacher who conducts observations full-time at many schools). Teachers' performance during classroom observations is scored using the district's own Teaching and Learning Framework (TLF) rubric.[2] Observers assign scores in several areas of practice that are averaged within observations, and then these composites are averaged across observations.[3]

[1] In the 2011-12 school year (and before) IVA was the only student learning component for Group 1 teachers even though these teachers do have TAS scores.
[2] The TLF rubric is modified somewhat for teachers in kindergarten and younger classrooms, and teachers who work with special education or English language learner students in non-traditional settings. During the period of our data, a separate rubric was used for teachers working with students with autism.
[3] Examples of areas of practice include "explains content clearly", "engages students at all learning levels", "provides students multiple ways to move toward mastery", "checks for student understanding", "maximizes instructional time and builds a supportive", and "learning-focused classroom."

The two remaining components are assessed by school administrators. Principals rate teachers' "commitment to the school community" (CSC) using a rubric that covers partnerships with parents, collaboration with colleagues, and support for school-wide initiatives and high expectations. Last, the school principal can deduct points from a teacher's final IMPACT score on the basis of poor attendance, tardiness, or other failures of "core professionalism" (CP).

Teachers' final IMPACT scores are a weighted average of the various components; these weights (summarized in Appendix Table D1) changed between the school years 2011-12 and 2012-13. IMPACT scores determine teachers' rating categories, based on pre-specified ranges. Teachers in the "ineffective" category are dismissed, as are teachers who fall in the "minimally effective" category for two consecutive years. At the other end of the distribution, teachers scoring in the "highly effective" category receive a one-time bonus of as much as $25,000. If a teacher is rated highly effective for two consecutive years, she receives a substantial permanent increase in salary; Dee and Wyckoff (2015) estimate this could be worth as much as a 29 percent increase in the present value of a teacher's total earnings over a 15-year horizon.

**References**

Dee, Thomas, and James Wyckoff. 2015. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT." *Journal of Policy Analysis and Management* 34(2): 267-297.

Appendix Table D1—IMPACT component weights

| | 2011-12 | | 2012-13 and 2013-14 | | 2014-15 and 2015-16 | |
|---|---|---|---|---|---|---|
| | Group 1 | Group 2 | Group 1 | Group 2 | Group 1 | Group 2 |
| Individual value-added | 0.50 | | 0.35 | | | |
| Teacher assessed student learning | | 0.10 | 0.15 | 0.15 | 0.15 | 0.15 |
| Teaching and learning framework | 0.35 | 0.75 | 0.40 | 0.75 | 0.75 | 0.75 |
| Commitment to school community | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 |
| School value-added | 0.05 | 0.05 | | | | |

**Appendix E: Transformation of IVA from DCPS 1.0 – 4.0 Scale**

DCPS provided us with the IVA ratings on a scale from 1.0 to 4.0, which it uses to calculate teachers' IMPACT scores and ratings. The scaled scores are generated by (1) assigning a score of 1.0 to teachers below the a certain cutoff and a score of 4.0 to those above a certain cutoff; (2) assigning the median teacher a score of 3.0; (3) assigning scores in the set (1.0 3.0) to teachers above the lower cutoff and below the mean using linear projection and rounding to one decimal place; and (4) assigning scores in the set (3.0 4.0) to teachers above the average but below the top cutoff using linear projection and rounding to one decimal place. The lower (upper) cutoff is generally around the $10^{th}$ ($90^{th}$) percentile, but vary by subject and year. Appendix Figure E1 shows histograms of these "IMPACT IVA" measures, which do not resemble the normal distribution typical of value-added estimates, are highly skewed due to "centering" at 3.0 (rather than 2.5), and are truncated at the tails.

To align the performance metric as closely as possible with the underlying value-added estimates, we use the assumption that IVA is normally distributed and "invert" the CDF of the IMPACT scaled IVA measures to generate IVA measures, with the center of the distribution corresponding to 3.0 and the tails assigned the average value conditional on being above/below the extreme value.[1] The distribution of these "transformed" IVA measures align better with our priors, but still have truncation in the tails (see Figure E2).

Fortunately, we acquired raw (i.e. not rescaled) IVA estimates for 2011, and can test how well our transformed IVA measures match up. The correlation with the raw IVA and the scaled measures is close to 0.95. The correlation between our transformed IVA and raw IVA at the individual level is about 0.98, with disagreement coming only in the extreme, truncated values

[1] For example, consider teachers who received a score of 4.0 in 2012 for reading. The cutoff to receive a 4.0 was the $86.4^{th}$ percentile. For these 13.6 percent of the sample we assign a value-added score of 1.837.

(see Figure E3). If we take the *average* raw IVA at each scaled IVA level and compare it with our transformed variable, we find a correlation of close to 0.999 in both subjects.

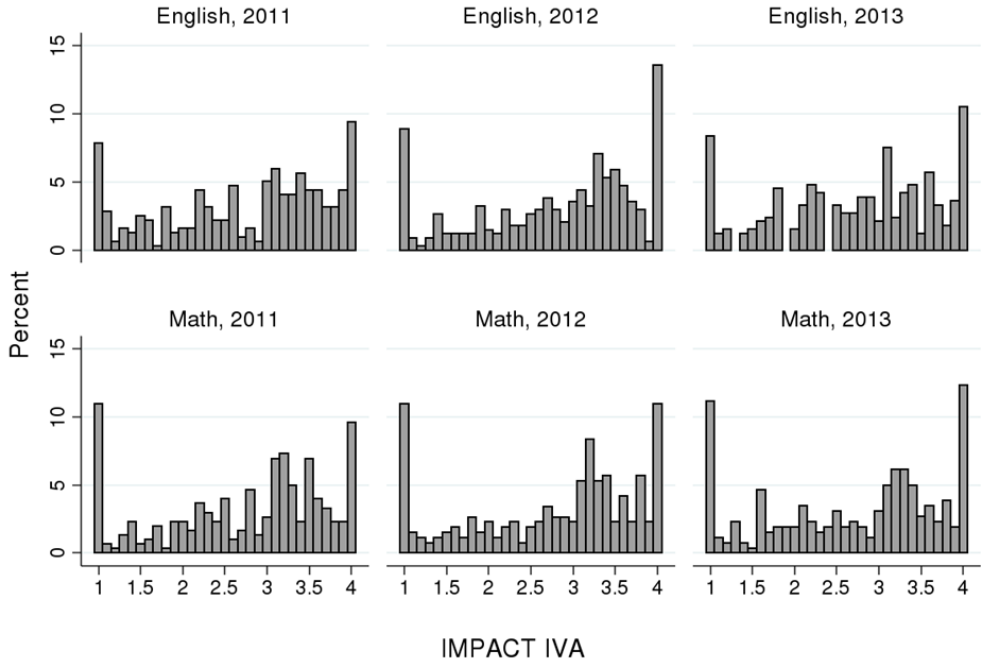



Figure E1: Distribution of Value-Added Measures on the DCPS Scale

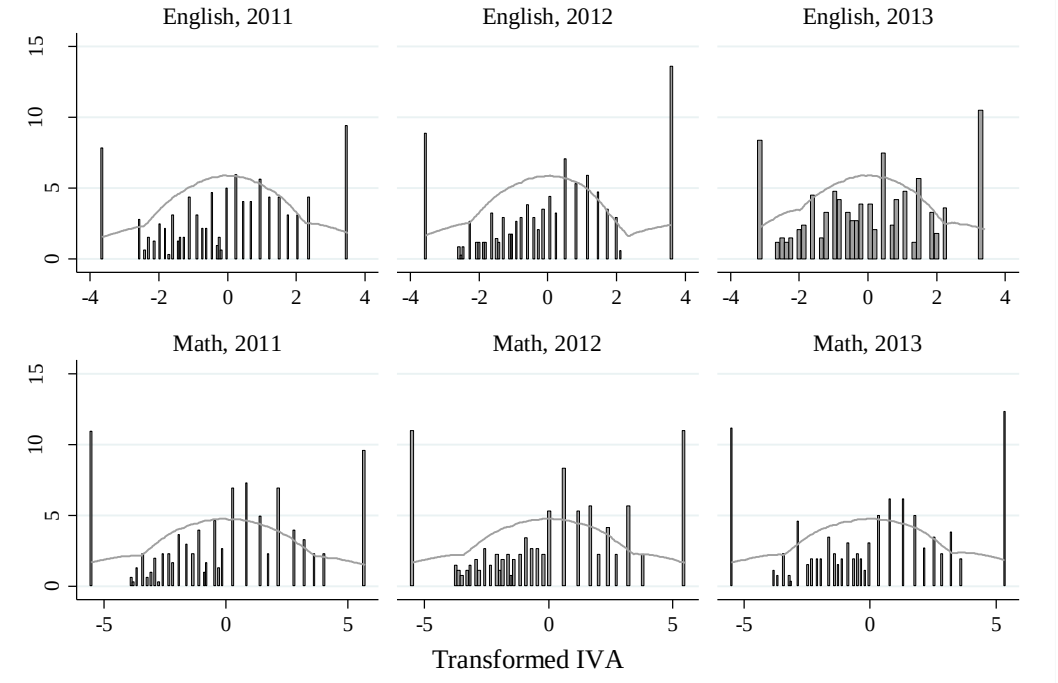



Figure E2: Distribution of Transformed Value-Added Measures

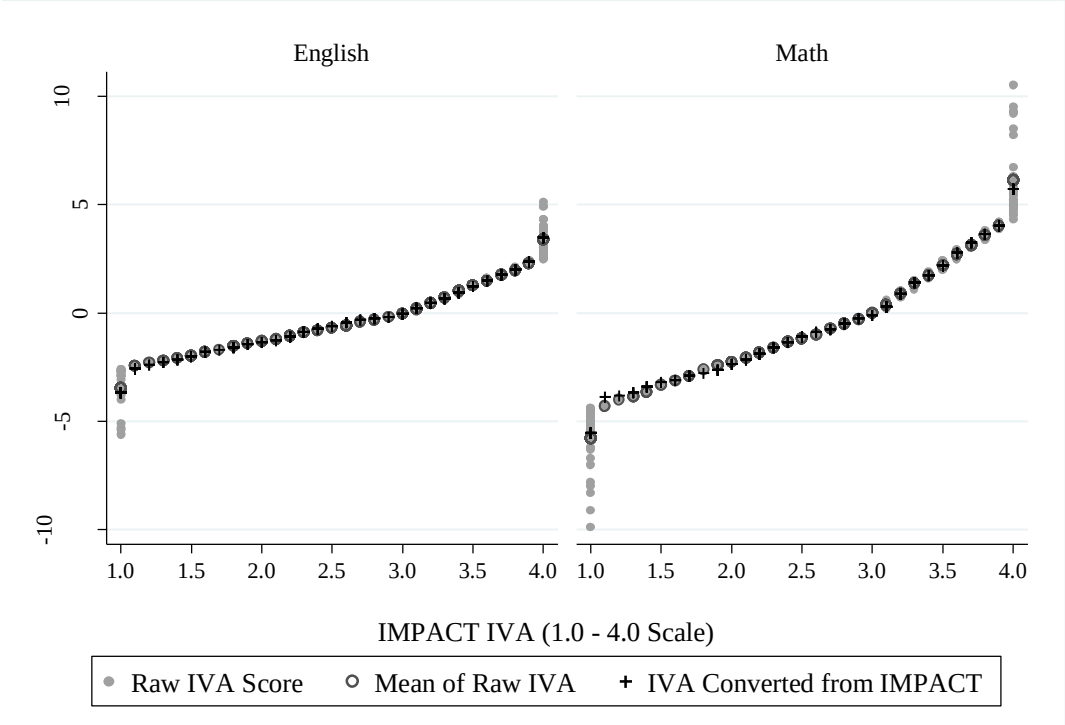



Figure E3: Comparison of Raw and Transformed Value-Added Measures in 2011

## Appendix F: Sample and Data Creation

We use data on over 7,000 individuals who applied through TeachDC in the years 2011-2013 and who were eligible for a teaching license in DC. We drop 198 applicants who participated in a Fast Track application option in 2011 and for whom some data were not collected. Our results are not sensitive to including these applicants. We analyze subsequent hiring and performance data from the school years 2011-12 through 2015-16.

We focus on applicants who applied for teaching jobs, and among those applicants identify new hires who are working as a DCPS teacher (as opposed to working as a counselor or administrator or any other role). We define a DCPS teacher as someone who (i) held a teaching position, as recorded in district human resources data and identified by union status, (ii) at some point during the school year. This definition includes individuals who were hired, worked in the fall, but left midyear. It also includes individuals hired midyear. Part-year teachers sometimes do not have job performance data (IMPACT scores), but we nevertheless count them as new hires. Additionally, some DCPS employees who are not officially holding a teaching position do have teaching responsibilities, and are scored in the IMPACT teacher performance evaluation program. In addition to the definition above, we count anyone with IMPACT teaching scores as a DCPS teacher. There are only two such teachers among our applicants, and the results are not sensitive to excluding them.

To obtain the PCK score we first standardize (mean zero, standard deviation one within years) the subject-specific essay scores on content and writing quality, as well as the KMT score. Our "PCK score" is the average of all standardized scores available for a teacher. For 2011 and 2012 applicants, our "interview score" is the average of two component scores, each standardized: (a) the mean of the applicant's TLF scores for the mini-lesson, and (b) the mean of

the applicant's behavioral interview questions scores. For 2013 applicants, we do not have separate scores for the mini-lesson and interview questions, but we have scores on several components (e.g., "instructional expertise," "communication skills") as well as several binary judgments (i.e., "outstanding," "no reservations," "reservations") which we combine using factor analysis to create the 2013 interview score. For each of the three years, a factor analysis on the components of the audition score yields just one factor, and we use the factor analysis weights in each year to construct our audition score. We get virtually identical results if we use a simple unweighted average of the component scores within the audition measure.