EdWorkingPaper No. 22-648

# Can a Commercial Screening Tool Help Select Better Teachers?

Olivia L. Chi
Boston University

Matthew A. Lenard
Harvard University

Improving teacher selection is an important strategy for strengthening the quality of the teacher workforce. As districts adopt commercial teacher screening tools, evidence is needed to understand these tools' predictive validity. We examine the relationship between Frontline Education's TeacherFit instrument and newly hired teachers' outcomes. We find that a one SD increase on an index of TeacherFit scores is associated with a 0.06 SD increase in evaluation scores. However, we also find evidence that teachers with higher TeacherFit scores are more likely to leave their hiring schools the following year. Our results suggest that TeacherFit is not necessarily a substitute for more rigorous screening processes that are conducted by human resources officials, such as those documented in recent studies.

VERSION: September 2022

Suggested citation: Chi, Olivia L., and Matthew A. Lenard. (2022). Can a Commercial Screening Tool Help Select Better Teachers?. (EdWorkingPaper: 22-648). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/6a2r-8j08

# Can a Commercial Screening Tool Help Select Better Teachers?

Olivia L. Chi
*Boston University*

Matthew A. Lenard
*Harvard University*

September 2022

## Abstract

Improving teacher selection is an important strategy for strengthening the quality of the teacher workforce. As districts adopt commercial teacher screening tools, evidence is needed to understand these tools' predictive validity. We examine the relationship between Frontline Education's TeacherFit instrument and newly hired teachers' outcomes. We find that a one SD increase on an index of TeacherFit scores is associated with a 0.06 SD increase in evaluation scores. However, we also find evidence that teachers with higher TeacherFit scores are more likely to leave their hiring schools the following year. Our results suggest that TeacherFit is not necessarily a substitute for more rigorous screening processes that are conducted by human resources officials, such as those documented in recent studies.

**Keywords:** school/teacher effectiveness, teacher characteristics, econometric analysis, educational policy

We thank Martin West, Eric Taylor, David Deming, seminar participants at Harvard and Boston University, conference participants at AEFP, as well as the editors and anonymous reviewers for their valuable comments and feedback. We are especially grateful to the Wake County Public School System for providing data. The research reported here was supported, in part, by the Institute of Education Sciences, U.S. Department of Education, through grant R305B150010 to Harvard University. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. Additional support came from the Multidisciplinary Program in Inequality and Social Policy at the Harvard Kennedy School. All remaining errors are our own.

# Introduction

An extensive literature documents substantial variation in teacher quality, and evidence shows that teacher quality impacts students' long-run outcomes (see Jackson, Rockoff, & Staiger, 2014 for a review). An important strategy for strengthening the quality of the teacher workforce is to identify and hire the applicants who will be most effective in the classroom. Recent studies suggest that information gathered during a district's application-screening process could be compiled to make predictions about teacher effectiveness. Goldhaber et al. (2017) find that scores from teacher selection rubrics used to rate applicants in Spokane Public Schools predict teacher value-added to test scores and teacher retention. Using data from Washington, DC, Jacob et al. (2018) find that applicants' background measures (e.g., undergraduate GPA) and scores on screening measures are predictive of teachers' evaluation scores. Bruno and Strunk (2019) use data from the Los Angeles Unified School District to show that scores from the district office's standardized screening system are predictive of teacher impacts on test scores, evaluation scores, and attendance. Sajjadiani et al. (2019), applying machine learning techniques to data from the Minneapolis Public School District, find that the relevance of applicants' work experience and their attributions for leaving past jobs predict teacher performance (student evaluations, observation scores, and value-added to test scores) and turnover.

These recent studies demonstrate promising ways in which information from the time of application can be harnessed to select better teachers. However, in these studies, districts' central office human resources officials play a large role in prescreening and scoring applications. This systematic collection and scoring of applicant information for use in hiring decisions may be a barrier in some school districts. Prior research suggests that district central offices vary in how they support schools in recruiting and screening teacher applicants, and principals vary in the degree to which they use data in their hiring decisions (Cannata et al., 2017). The teacher hiring process may even be rushed and information-poor (Liu & Johnson, 2006).

To overcome this limitation, school districts across the United States are turning to the private sector for new teacher screening tools that claim to use “big data” to help identify better teachers from the pool of teacher applicants (Flanigan, 2016; Simon, 2014). Commercial screening instruments have existed for decades, such as the Haberman Star Teacher PreScreener and Gallup’s Teacher Perceiver Interview, with limited evidence suggesting that scores from these two tools are modestly related to teacher performance (Metzger & Wu, 2008; Rockoff et al., 2011; see Appendix A for additional background.) However, additional commercial screening instruments have arrived on the market, boasting data-driven screening scores (Simon, 2014). Many districts currently pay firms to use these screening instruments, which typically include assessments that applicants take while completing online teacher job applications (Simon, 2014). Despite the growing popularity of such commercial screening tools, there exists limited evidence on whether these tools are effective in predicting teacher performance.

To extend this literature, we study the extent to which applicants’ scores from a “big data” commercial screening tool, Frontline Education’s TeacherFit instrument, predict teacher outcomes in a large U.S. school district. Specifically, we ask: Are the results from the TeacherFit screening tool predictive of teachers’ evaluation scores, absences, retention in the same school/district, and impacts on student test scores?

We use data from the Wake County Public School System (WCPSS) in North Carolina, the 14th largest school district in the nation (de Bray, 2021). Beginning in January 2016, WCPSS required teacher applicants to take the TeacherFit assessment to submit their applications. We study the predictive validity of applicants’ TeacherFit scores among the new teacher applicant pool for WCPSS in school years 2016-17 and 2017-18. We find that a one standard deviation increase on an index of TeacherFit scores is associated with a 0.06 standard deviation increase in the evaluation scores that teachers receive from principals. In addition, we find evidence that teachers with higher TeacherFit scores are more likely to leave their hiring schools after the first year. We do not find a significant relationship between TeacherFit scores and either value-added to math or ELA test scores. To alleviate concerns of bias from sample selection, we estimate a Heckman selection model and find that sample-selection corrected estimates are similar to our estimates without corrections.

## Setting, Data, and Measures

Frontline Education is a school administration software provider whose broad portfolio of products reaches over 10,000 clients (Frontline Education, 2021). WCPSS began using Frontline's web-based recruiting platform, Frontline Recruiting and Hiring, to manage job postings and applications in the summer of 2015. Beginning in January 2016, WCPSS required teacher applicants to take Frontline's screening assessment, TeacherFit, to submit their applications via the district's site on Frontline's online platform. Frontline Education's website states that their assessments are "driven by university backed research" and "leverage 'machine learning'" (Grunwell, 2016). According to a sales webinar, the assessment creators developed the items based on interviews with subject matter experts, analysis of job descriptions, and reviews of research, followed by testing of the items on teacher and school employees (Reese, 2018). "Hundreds of thousands" of applicants have taken the assessment since it became available in 2008 (Frontline Education, 2022). However, to our knowledge, the extent to which scores from the TeacherFit instrument can identify effective teachers has not yet been documented in peer-reviewed literature.

The TeacherFit assessment, which claims to help “identify outstanding teachers” (Grunwell, 2016), takes approximately 20-30 minutes, does not allow for blank responses, and does not have obvious “correct” or “incorrect” answers. Rather, the items attempt to assess applicants’ attitudes, beliefs, habits, and personality traits by requiring applicants to address situational prompts and attitudinal statements by selecting Likert scale responses. Following the assessment, Frontline constructs the applicants’ scores and makes them available to WCPSS administrators and school principals on the online hiring dashboard. Each candidate receives an overall score and separate scores on each of the 6 dimensions: Fairness & Respect, Concern for Student Learning, Adaptability, Communication & Persuasion, Planning & Organizing, and Cultural Competence. Each score falls on a 1-to-9 scale. (See Appendix A for additional background on TeacherFit.)

While applicants’ scores were made available to administrators on the hiring platform, school principals did not receive strict or explicit guidance from the WCPSS Human Resources office on how to use or interpret the scores, keeping with a tradition of a decentralized hiring process in WCPSS. While principals received communication that TeacherFit scores could help identify strong candidates, they also received messaging that the TeacherFit assessment is only one part of the hiring process, and they were free to pursue candidates that do not score well. (See Appendix B for additional background on teacher hiring.)

We use application and administrative data to study the new teacher applicant pool for school years 2016-17 and 2017-18. We link teacher applicants’ scores and application information to WCPSS administrative data, which includes teacher characteristics, assignments, evaluation scores, absences, and links to students. Individuals who are not previously observed as WCPSS teachers are included in the new teacher applicant pool if they: 1) submitted teacher applications in the calendar years 2016 and 2017 and/or 2) are newly hired teachers in 2016-17 and 2017-18. The applicant pool includes 12,548 individuals, of whom 2,367 are observed as newly hired teachers in either 2016-17 or 2017-18. However, TeacherFit scores are missing for 8% of the applicant pool. Therefore, 11,491 individuals, of whom 2,104 are observed as new hires across 184 schools, are eligible for inclusion in the analyses below.

Table 1 provides summary statistics of the TeacherFit scores for the individuals in the teacher applicant pool with non-missing scores (N = 11,491). The TeacherFit overall score (1-9 scale) has a mean of 6.04 (SD = 1.80), while the means of the scores on the 6 dimensions range from 5.23 (SD = 2.08) in Cultural Competence to 6.12 (SD = 1.82) in Adaptability. In the analyses below, we use a TeacherFit index score, which we construct by summing the scores for the 6 dimensions and then standardizing to have a mean of 0 and unit standard deviation.

## Teachers' Outcome Measures

Our primary outcome of interest is teachers' evaluation scores. In the North Carolina Teacher Evaluation Process, teachers must be reviewed annually by their principals or a similar designated evaluator. To construct teachers' annual evaluation scores, we fit a Graded Response Model (GRM) on the ratings that teachers receive on each element of their Summary Rating Forms from the North Carolina evaluation regime (Kraft et al., 2020). GRM models are in the family of Item Response Theory (IRT) models that are commonly used in educational and psychological assessment. GRMs are developed for ordered categorical items, such as the 5-category scale on the Summary Rating Forms in the NC Teacher Evaluation Process (Samejima, 1968). We then standardize the GRM scores within-year to have a mean of 0 and a standard deviation of 1.

We also examine additional teacher outcome measures, including the number of days a teacher is absent, retention in the same school and same district in the following year, and impacts on math and ELA test scores, when available. Retention in the same school (district) is a binary indicator of whether a teacher returns to teach in the same school (in WCPSS) in the following school year. By definition, those who remain teaching in the same school in the following year also remain teaching in the same district, WCPSS, in the following year. However, those who remain teaching in WCPSS in the following year are not necessarily teaching in the same school in the following year. To calculate teachers' impacts on test scores, we estimate value-added models for math and ELA teachers of 4th through 8th grade students (see Appendix C).

## Empirical Strategy

To estimate whether TeacherFit scores are predictive of teacher-level outcomes, we use OLS to estimate:

$$Y_{jkt} = \delta_0 + \delta_1 \text{Score}_j \; + T_j\delta_2 + S_{jkt}\delta_3 + \delta_t + \delta_h + \nu_{jkt}, \qquad (1)$$

where $Y_{jkt}$ indicates the outcome of teacher $j$ in school $k$ at time $t$. $\text{Score}_j$ refers to teacher $j$'s standardized TeacherFit index scores. The coefficient of interest $\delta_1$ is the expected change in the outcome $Y$ associated with a one standard deviation increase in an applicant's TeacherFit index score. $T_j$ represents a vector of indicators for the teacher characteristics of race, gender, and experience. In theory, including controls for teacher experience may account for variation in the outcome that would instead be attributable to differences in TeacherFit scores in the absence of experience controls. However, we include these to address whether and to what extent TeacherFit scores can provide additional predictive information, above and beyond what is already known from resumes at the time of hiring, such as teacher experience.

$S_{jkt}$ represents a vector of annual school characteristics of teacher $j$'s school $k$, including student gender, race, Limited English Proficient (LEP) status, and special education status, aggregated to school-level, along with mean school-level prior test scores and school size. $\delta_t$ represents year indicators, and $\delta_h$ are indicators for the number of years since being newly hired. Standard errors are clustered at the teacher level.

To alleviate concerns of bias stemming from the possibility that teachers with higher (or lower) TeacherFit scores are systematically sorting into schools or job assignments (e.g., 3rd grade, middle/secondary math, middle/secondary ELA) that enable teachers to have better performance measures, we also estimate these models with: 1) school fixed effects in place of school-level characteristics, and 2) job assignment fixed effects. In these models, the identifying variation comes from applicants that are hired into the same school and applicants that are hired into the same job assignment.

A limitation worth noting is the potential for bias stemming from the possibility that administrators may - subconsciously or otherwise - reward higher TeacherFit scores, which they observed during the hiring process, with higher evaluation scores. This could bias our estimate of the relationship between TeacherFit and evaluation scores upwards. However, we suspect that this would not be a large source of bias in this specific context as administrators were not advised or guided to put much stock in TeacherFit scores.

## Sample Selection Correction

While we can observe the outcomes of interest for hired applicants, we lack information on how non-hired applicants would have performed had they been hired. In other words, we cannot examine the relationship between TeacherFit scores and teacher performance for the full range of applicants. Given that the individuals making hiring decisions in WCPSS are likely trying to select applicants that they perceive to be of higher quality, the hiring process may introduce selection bias into our estimates. Specifically, we are concerned that low scoring individuals who end up hired as new teachers in WCPSS, in spite of their low scores, are particularly impressive in ways that are: a) unobservable and b) correlated with their performance as teachers. To alleviate concerns of bias from sample selection, similar to Goldhaber et al. (2017), we estimate sample selection-corrected models using a Heckman selection model (Heckman, 1979). We identify the model using a function of the school size growth among the set of schools to which applicants submit applications (see Appendix D). The extent of school size growth among an applicant's set of schools is predictive of the likelihood they become a newly hired teacher in WCPSS but is unrelated to on-the-job performance as a teacher. We find that sample-selection corrected estimates are similar to our estimates without corrections, alleviating concerns of large bias from sample selection.

## Results

We present estimates from equation (1) in Table 2. The odd-numbered columns include controls for school characteristics, while the even-numbered columns replace the school characteristics controls with school fixed effects and add job assignment fixed effects. As shown in Columns (1) and (2), we find that scores from the TeacherFit screening tool significantly predict teachers' standardized evaluation scores. A one standard deviation increase in the

TeacherFit index is associated with a 0.08 SD increase in evaluation scores in our baseline model. After including school and job assignment fixed effects, this estimate attenuates slightly but remains statistically significant at 0.06 SD. This magnitude is about 16% of the estimated within-teacher returns to experience after 1 year of teaching (0.38 SD, estimated with WCPSS data from 2015-16 through 2017-18). Columns (3) and (4) examine the relationship between the TeacherFit scores and teacher absences. The coefficients are positive and small, but the results are not statistically significant—null results that are consistent with those reported by both Goldhaber et al. (2017) and Rockoff et al. (2011).

In columns (5) through (8), we examine the relationship between TeacherFit scores and teacher retention. Surprisingly, in our baseline models, we find that a one standard deviation increase in the Teacher Fit index is associated with a 3.4 percentage point *decrease* in the likelihood of remaining as a teacher in the same school in the following year (column (5)) and a 2.4 percentage point *decrease* in the likelihood of remaining in WCPSS in the following year (column (7)). After including school and job assignment fixed effects, the results attenuate slightly. These retention results provide some evidence that TeacherFit scores are negatively associated with within-school and within-district teacher retention.

These findings are contrary to the retention results from Goldhaber et al. (2017), who find that higher scores on a screening rubric, which is completed by human resources hiring officials, predict an increase in district retention, as well as results from Jacob et al. (2018), who find that higher screening scores predict a higher likelihood of remaining in the hiring school. Jacob et al. (2018), however, do find that teachers with better academic background scores are more likely to leave their hiring school and more likely to leave DCPS after their first year.

Columns (9) through (12) present results of the relationship between TeacherFit scores and value-added to math and ELA test scores of students in grades 4 through 8, measured in student-level test score SDs. All our estimates are relatively close to 0 and are statistically insignificant. The point estimates for math value-added scores, 0.005 (column 9) and 0.004 (column 10) are equivalent to 0.025 SD and 0.020 SD, respectively, in teacher-level SDs. The point estimates for ELA value-added scores, 0.002 (column 11) and -0.002 (column 12) are equivalent to 0.014 SD and -0.014 SD, respectively, in teacher-level SDs.

### Selection-corrected Estimates

Table 3 presents our selection-corrected estimates of the relationship between TeacherFit scores and our outcomes of interest. Here, the sample of hired teachers is smaller than that included in Table 2, as it is limited to applicants who: a) submitted teacher applications in the same calendar year in which they are hired, and b) apply to schools where we can measure the change in school size between the prior year and the time of application. We also only include newly hired teachers' first-year in the data, and the model is fit at the person-level. In this table, we present estimates from our baseline model (i.e., including school characteristics, and absent school and job assignment fixed effects), displaying results without and with the sample correction in the odd and even columns, respectively. The magnitude of the estimates appear substantively similar without and with the sample selection correction, alleviating concerns of large bias introduced by sample selection. Given these results, we prefer the estimates presented in Table 2 from our larger and less restrictive sample.

## Conclusion

We find that scores from the TeacherFit instrument have some capacity to predict teacher performance as measured by evaluation scores from principals. We find that a one standard deviation increase on an index of TeacherFit scores is associated with a 0.06 standard deviation increase in evaluation scores. However, we do not find a significant relationship between TeacherFit scores and teacher impacts on test scores. Furthermore, we find some evidence that teachers with higher TeacherFit scores are more likely to leave their hiring schools after the first year.

These results suggest that the TeacherFit commercial screening tool is not necessarily a substitute for the promising screening processes that are conducted by human resources officials as described in the studies by Bruno and Strunk (2019), Goldhaber et al. (2017), and Jacob et al. (2018). The screening scores from these more elaborate screening processes appear to be stronger predictors of desirable teacher outcomes, and investing in these screening systems may have higher payoffs than investing in commercial screening tools that may be cheaper and easier to implement. These TeacherFit results are also more modest than those documented in Rockoff et al.'s (2011) examination of Haberman PreScreener scores, though the differences in teacher characteristics between studies (WCPSS new hires vs. New York City elementary/middle math teachers) make comparison difficult. Nevertheless, our results focus on just one example of a "big data" commercial screening instrument. As districts adopt and/or continue using commercial screening tools, researchers and practitioners should monitor the predictive validity of these tools to ensure that scores from these tools contain information that can be used to improve teacher selection and retention.

## References

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

Cannata, M., Rubin, M., Goldring, E., Grissom, J. A., Neumerski, C. M., Drake, T. A., & Schuermann, P. (2017). Using teacher effectiveness data for information-rich hiring. *Educational Administration Quarterly*, *53*(2), 180–222.

de Brey, C., Synder, T. D., Zhang, A., & Dillow, S. A. (2021). Digest of Education Statistics 2019 (NCES 2021-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Flanigan, R. L. (2016, January 25). More Districts Mine Data to Refine Teacher Hiring. *Education Week*. https://www.edweek.org/

Frontline Education. (2021, December 27). *Frontline Education recognized for adding value to K-12 administrators and the edtech industry by a variety of organizations in 2021* [Press release]. Retrieved from https://www.frontlineeducation.com/news/frontline-education%E2%80%AFrecognized-for-adding%E2%80%AFvalue%E2%80%AFto-k-12-administrators-and-the-edtech-industry-by-a-variety-of-organizations-in-2021/

Frontline Education (2022). Application Screening Assessments. Retrieved June 28, 2022, from https://www.frontlineeducation.com/solutions/recruiting-hiring/screening-assessments/

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

Grunwell, A. (2016, August 1). *Applicant Screening Assessments: Frequently Asked Questions*. Applicant Screening Assessments: Frequently Asked Questions. https://www.frontlineeducation.com/blog/applicant-screening-assessments-faqs/

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153. https://doi.org/10.2307/1912352

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher Effects and Teacher-Related Policies. *Annual Review of Economics*, *6*(1), 801–825. https://doi.org/10.1146/annurev-economics-080213-040845

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

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Kraft, M. A., Papay, J. P., & Chi, O. L. (2020). Teacher skill development: Evidence from performance ratings by principals. *Journal of Policy Analysis and Management*, *39*(2), 315-347. https://doi.org/10.1002/pam.22193

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

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

Reese, S. (2018). *Applitrack Selection Assessments* [Webinar]. Frontline Technologies. https://www.youtube.com/watch?v=Tl7ngKG1xoI

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

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

Samejima, F. (1968). ESTIMATION OF LATENT ABILITY USING A RESPONSE PATTERN OF GRADED SCORES. *ETS Research Bulletin Series*, *1968*(1), i–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Simon, S. (2014, December 29). Teacher hopefuls go through big data wringer. *Politico*. https://www.politico.com

## Tables

Table 1: Summary Statistics of Scores for the New Teacher Applicant Pool

| | New Teacher Applicant Pool | | | | New Hires | | | | Non-hires | | | | Difference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Min. | Max. | Mean | SD | Min. | Max. | Mean | SD | Min. | Max. | New Hires - Non-hires |
| TeacherFit Overall Score | 6.04 | 1.80 | 1.00 | 9.00 | 6.21 | 1.69 | 1.00 | 9.00 | 6.01 | 1.82 | 1.00 | 9.00 | 0.21*** |
| *Dimensions* | | | | | | | | | | | | | |
| Fairness and Respect | 5.56 | 1.84 | 1.00 | 9.00 | 5.79 | 1.75 | 1.00 | 9.00 | 5.51 | 1.86 | 1.00 | 9.00 | 0.28*** |
| Concern for Student Learning | 5.95 | 1.86 | 1.00 | 9.00 | 6.17 | 1.74 | 1.00 | 9.00 | 5.91 | 1.88 | 1.00 | 9.00 | 0.26*** |
| Adaptability | 6.12 | 1.82 | 1.00 | 9.00 | 6.13 | 1.73 | 1.00 | 9.00 | 6.11 | 1.84 | 1.00 | 9.00 | 0.01 |
| Communication and Persuasion | 6.05 | 1.85 | 1.00 | 9.00 | 6.06 | 1.80 | 1.00 | 9.00 | 6.05 | 1.86 | 1.00 | 9.00 | 0.01 |
| Planning and Organizing | 6.11 | 1.83 | 1.00 | 9.00 | 6.23 | 1.75 | 1.00 | 9.00 | 6.09 | 1.84 | 1.00 | 9.00 | 0.15** |
| Cultural Competence | 5.23 | 2.08 | 1.00 | 9.00 | 5.48 | 2.01 | 1.00 | 9.00 | 5.17 | 2.09 | 1.00 | 9.00 | 0.30*** |
| *Composite Measures* | | | | | | | | | | | | | |
| TeacherFit Sum Score (raw) | 35.03 | 8.48 | 6.00 | 54.00 | 35.85 | 8.02 | 7.00 | 54.00 | 34.84 | 8.57 | 6.00 | 54.00 | 1.01*** |
| Std. TeacherFit Sum Score | 0.07 | 0.98 | -3.27 | 2.26 | 0.17 | 0.92 | -3.16 | 2.26 | 0.05 | 0.99 | -3.27 | 2.26 | 0.12*** |
| | | | | | | | | | | | | | |
| Missing TeacherFit Score | 0.08 | | | | 0.11 | | | | 0.08 | | | | 0.03*** |
| N (applicants with non-missing scores) | 11,491 | | | | 2,104 | | | | 9,387 | | | | 11,491 |
| N (teachers) | 12,548 | | | | 2,367 | | | | 10,181 | | | | 12,548 |

Notes: ** p<0.01, *** p<0.001. The new teacher applicant pool includes 12,548 individuals who submitted teacher applications in the calendar years 2016 and 2017 and/or are newly hired teachers in 2016-17 and 2017-18.

Table 2: Relationship between TeacherFit Scores and Teacher Outcomes

| | Evaluation Scores | | Absences | | Stay in School | | Stay in District | | Math VA | | ELA VA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Std. TeacherFit Sum Score | 0.077*** | 0.060*** | 0.059 | 0.226 | -0.034*** | -0.029** | -0.024** | -0.021* | 0.005 | 0.004 | 0.002 | -0.002 |
| | (0.019) | (0.017) | (0.166) | (0.171) | (0.009) | (0.010) | (0.008) | (0.009) | (0.016) | (0.019) | (0.011) | (0.013) |
| School Characteristics | Y | | Y | | Y | | Y | | Y | | Y | |
| School FE | | Y | | Y | | Y | | Y | | Y | | Y |
| Job Assignment FE | | Y | | Y | | Y | | Y | | Y | | Y |
| Unique Teachers | 1,884 | 1,884 | 2,063 | 2,063 | 2,088 | 2,088 | 2,088 | 2,088 | 236 | 236 | 246 | 246 |
| N (teacher-years) | 2,580 | 2,580 | 2,830 | 2,830 | 2,873 | 2,873 | 2,873 | 2,873 | 302 | 302 | 316 | 316 |
| N (teacher-years) w/ 2+ teachers | | | | | | | | | | | | |
| in same school | | 2,578 | | 2,828 | | 2,871 | | 2,871 | | 282 | | 292 |
| in same job assignment | | 2,580 | | 2,828 | | 2,871 | | 2,871 | | 302 | | 316 |

Notes: * p<0.05, ** p<0.01, *** p<0.001. Clustered standard errors at the teacher-level are in parentheses. All regressions include a vector of indicators for the teacher characteristics of race, gender, and experience. School characteristics include student gender, race, Limited English Proficient (LEP) status, and special education status, aggregated to school-level, along with mean school-level prior test scores and school size.

Table 3: Sample Selection Corrected Estimates

| | Evaluation Score | | Absences | | Stay in School | | Stay in District | | Math VA | | ELA VA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
| Std. TeacherFit Sum Score | 0.079*** | 0.091** | 0.204 | 0.209 | -0.028* | -0.020 | -0.027** | -0.042** | 0.028+ | 0.018 | 0.015 | 0.027+ |
| | (0.020) | (0.028) | (0.185) | (0.262) | (0.012) | (0.017) | (0.010) | (0.015) | (0.016) | (0.018) | (0.014) | (0.015) |
| Mills ratio | | 0.231 | | 0.084 | | 0.137 | | -0.246 | | -0.214 | | 0.250 |
| | | (0.382) | | (3.382) | | (0.211) | | (0.185) | | (0.198) | | (0.154) |
| Sample selection correction | | Y | | Y | | Y | | Y | | Y | | Y |
| N (teachers) | 1,662 | 1,662 | 1,844 | 1,844 | 1,874 | 1,874 | 1,874 | 1,874 | 232 | 232 | 231 | 231 |
| N (applicants first-stage) | | 10,706 | | 10,701 | | 10,701 | | 10,701 | | 10,713 | | 10,714 |
| F (excluded variable) | | 23.845 | | 22.442 | | 25.840 | | 25.840 | | 16.891 | | 13.645 |

Notes: + p<0.10, * p<0.05, ** p<0.01, *** p<0.001. Clustered standard errors at the teacher-level are in parentheses. All regressions include a vector of indicators for the teacher characteristics of race, gender, and experience, as well as student gender, race, Limited English Proficient (LEP) status, and special education status, aggregated to school-level, along with mean school-level prior test scores and school size.

# Appendix

## Appendix A: Background on TeacherFit

According to their website, Frontline Recruiting and Hiring offer 5 assessments, which are based on research conducted by John Arnold, Ph.D., of Wayne State University, and Neal Schmitt, Ph.D., of Michigan State University. They claim that these assessments help "find the right candidate for every job type" (Frontline Education, 2022), and they describe their assessments as follows (quoted from their website):

- "**TeacherFit**: Identify outstanding teachers
- **TeacherFit SE**: Identify qualified special education teachers
- **TeacherFit Urban**: Identify great teachers for urban schools
- **JobFit**: Identify the best candidates for your support staff
- **AdminFit**: Identify winning administrators" (Frontline Education, 2022).

In this study, we examine the predictive validity of the TeacherFit assessment, which Frontline's promotional materials state has been available since 2008 and taken by hundreds of thousands of candidates in the United States (Frontline Education, 2022). As of 2015, when the product was known as Applitrack TeacherFit, the assessment was used in approximately 300 school districts (Applitrack, 2015, as cited in Oestreich, 2016) with nearly 170,000 candidates having completed the assessment (Frontline, 2015, as cited in McKinney, 2017).

According to an online webinar released by Frontline Technologies (Reese, 2018), TeacherFit's developers generated the assessment by: 1) creating a large bank of test items based on interviews with subject matter experts, analysis of job descriptions, and reviews of research; 2) administering the items to teachers and school employees identified as exemplary by their supervisors, and 3) conducting item analysis of the results to determine which items would be retained in the item bank.

To further understand TeacherFit items, we viewed an online demonstration version of the assessment from 2018. The demo version we accessed contained 1) "Situational" items and 2) "Attitude – Agree/Disagree" items (Applitrack, n.d.; Resse, 2018). The "Situational" items began by describing a school-based scenario that teachers might find themselves in. The applicant is then asked to respond to a series of items that represent potential follow-up actions to the stated scenario by indicating on a Likert scale how likely they are to take the listed actions. The "Attitude – Agree/Disagree" items present statements about attitudes, beliefs, habits, and personality traits, and the applicant is asked to use a Likert scale to indicate their level of (dis)agreement with the presented statement.

In our own view, it was not obvious how to answer the items to maximize one's score. While there were some items that one would likely assume that indicating agreement would help to increase one's score (e.g., whether one is often prepared, whether one can handle stressful situations), there were many items for which it was unclear how to best answer if one's goal were to obtain a high score. Similarly, it was not obvious how likely one should indicate they would take various actions in response to the presented scenarios to maximize one's score. In other words, we typically didn't find that there were obvious "right" and "wrong" answers to the assessments.

From an applicant's view, the characteristics of the TeacherFit assessment, which uses a web interface to present Likert-scale items related to teacher attitudes and beliefs, appear similar in format to other popular screening tools. For example, Gallup's TeacherInsight tool is a web-based assessment that takes approximately 40 minutes to complete, asking applicants to respond to statements using a Likert scale, complete multiple choice questions, and respond to open-ended questions to assess attitudes, beliefs, and behaviors (Metzger & Wu, 2008). Similarly, the Haberman Star Teacher PreScreener, an assessment intended to provide information about a candidate's effectiveness in an urban classroom, requires applicants to answer 50 multiple choice questions to assess traits such as persistence, organization and planning, beliefs about the value of students learning, and more (Haberman, 1993; Rockoff et al. 2011).

Though commercial screening tools by Gallup and Haberman have existed for decades,[1] there is limited peer-reviewed research on the efficacy of such screening tools. In a meta-analysis on the predictive validity of the predecessor of the TeacherInsight tool—Gallup's Teacher Perceiver Interview—Metzger and Wu (2008) find weak to moderate relationships with indicators of teacher outcomes, notably administrator ratings and teacher absences. However, it is important to note that their meta-analysis of 24 studies included only one peer-reviewed journal article (i.e., Young & Delli, 2002), while the remaining studies included doctoral dissertations and studies provided by Gallup. Rockoff et al. (2011) find positive relationships between scores on the Haberman Star Teacher PreScreener and both student and teacher outcomes. Specifically, they find that a one SD increase in the Haberman score is associated with a 0.02 SD increase in math achievement ($p = 0.07$) and a 0.16 SD increase in subjective evaluation by supervisors ($p = 0.02$) among survey respondents comprised of new elementary and middle school math teachers in New York City.

This limited literature suggests that these traditional commercial screening tools may have some potential to help identify effective teachers. However, in addition to further examining traditional commercial screening tools that have been on the market for years, school

[1] Gallup's Teacher Perceiver Interview was created in the 1970s (Metzger & Wu, 2008), and the Haberman Star Teacher PreScreener was developed in the 1980s (Rockoff et al., 2011).

districts may also benefit from the examination of newer commercial screening tools, such as TeacherFit, that claim to use “big data” or machine-learning techniques to improve teacher hiring (Grunwell, 2016; Simon, 2014). Little is known about the efficacy of TeacherFit, though two recent doctoral dissertations examine simple relationships between TeacherFit scores and teacher outcomes in smaller sized samples, finding weak to modest relationships (Oestreich, 2016; McKinney, 2017).

## Appendix B: Background on Teacher Hiring

The teacher selection process generally consists of 1) screening teacher applicants and identifying those who are most promising, 2) interviewing and/or gathering additional information about applicants via observations, demonstrations, sample work, references, etc., and 3) deciding who should receive a job offer (Johnson, 2020; Rose et al., 2014). School principals typically serve as the primary decision makers in the teacher hiring process (Cannata & Engel, 2012). A comprehensive view of the extent to which principals have autonomy in hiring in the U.S. is largely unavailable in the literature (Perrone & Meyers, 2021), though Engel et al. (2015) document increases in principals’ perceptions of their influence over the teacher hiring process between the late 1980s and 2012, particularly in urban districts.

The extent of involvement in the hiring process from staff members in district central offices can vary widely across school districts. Cannata et al. (2017) document substantial variation in how central offices screen teacher applicants prior to the continuation of the hiring process at the school site. For example, they describe two central offices that ask applicants only for their credentials. In contrast, another central office created a 5-stage process (of which 4 stages were conducted by central office staff), such that only applicants who passed each stage could be interviewed by school principals.

Cannata et al. (2017) also document variation in the extent to which central offices provide clear and consistent guidance about how to make their hiring decisions. Perhaps unsurprisingly, principals in systems that provided very little guidance also varied greatly in the processes and criteria that they used for hiring. Furthermore, the authors find that principals in settings without clear hiring expectations were less likely to use teacher effectiveness data to guide hiring decisions. These findings complement prior work by Liu and Johnson (2006), who document that the teacher hiring can be a rushed and information-poor process.

However, even in information-poor processes, principals are likely to make decisions based on some set of criteria. Prior research suggests that principals seek a variety of attitudes and dispositions from candidates, as well as a good “fit” between the values, preferences, and skills of teacher applicants and the organization [*person-organization fit*], the immediate work group [*person-group fit*], and the job position [*person-job fit*] (see Perrone & Meyers, 2021, for a review). The research literature also documents mixed findings about principals’ preferences for teachers’ professional characteristics, such as academic background (Perrone & Meyers, 2021), though two recent studies may be particularly relevant in shedding light on principals’ preferences for teacher characteristics during the screening process.

Hinrichs (2021) conducted a randomized field experiment, in which 6,000 fictitious resumes with randomly-selected characteristics were sent to schools, and finds that public schools respond more favorably to in-state applicants and candidates from more selective colleges. However, Hinrichs also finds that having a high GPA does not appear to be helpful in eliciting a favorable response from schools. In line with these findings, Jacob et al. (2018) find that “[f]or the most part, applicants’ academic credentials” (e.g., undergraduate GPA, SAT/ACT score) “appear to bear little relation to being hired” (p. 89). Moreover, Jacob and colleagues find that principals *did not* rely on the screening scores that were provided by the district in their hiring decisions. They do however find that local applicants are more likely to be hired, while those with no teaching experience were less likely to be hired.

Taken together, this existing literature provides some context for the hiring processes in WCPSS. During the time period of the study, central office staff members were not involved in screening applicants prior to schools conducting their own school-based hiring processes. Fitting with the tradition of decentralization in the hiring process in WCPSS, principals were not held to specific expectations about what criteria to use or weigh most heavily in the selection process. Principals generally had discretion in their hiring decisions, provided that their hires meet the minimum criteria for teaching in schools (e.g., holding credentials, passing background checks).

Without clear guidance about whether and how to use TeacherFit scores, we suspect—based on the findings discussed above, and in particular, those of Cannata et al. (2017) and Jacob et al. (2018)—that principals *did not* heavily weigh TeacherFit scores in their teaching hiring process, if they considered them at all. Rather, we suspect that principals used their own preferences to guide the criteria that they weighed most heavily in their decision making. This would comport with the descriptive statistics in Table 1, which illustrate that differences in TeacherFit scores between hires and non-hires in WCPSS were not very large. Our perceptions of the extent to which TeacherFit scores were considered are also aligned with those of district leaders, who have expressed that they believed principals were not using or paying attention to TeacherFit scores, and therefore discontinued the implementation of the TeacherFit assessment in more recent years.

## Appendix C: Value-added to Math and ELA test scores

To estimate teachers' impacts on test scores, we estimate a value-added model commonly used in the literature (Koedel et al., 2015):

$$A_{ijt} = f(A_{i,t-1}) + X_{ijt}\beta + \alpha_{jt} + \gamma_g + \epsilon_{ijt}, \tag{A1}$$

where $A_{ijt}$ is the outcome of student $i$ assigned to teacher $j$ in time $t$. We control for a function of prior math and ELA achievement, which includes the squared and cubed values of the same-subject prior test score. $X_{ijt}$ is a vector that includes student-, classroom-, and school-level covariates, including gender, race, Limited English Proficient (LEP) status, and special education status, as well as these covariates aggregated to the classroom- and school-level, along with mean classroom- and school-level prior test scores. $\gamma_g$ represent grade fixed effects. $\alpha_{jt}$ are teacher-by-year fixed effects and represent teacher $j$'s value-added to test scores in time $t$. We estimate the model separately for math and ELA teachers of 4th through 8th grade students.

## Appendix D: Sample Selection Correction

To alleviate concerns of bias from sample selection, similar to Goldhaber et al. (2017), we estimate sample selection-corrected models using a Heckman selection model (Heckman, 1979). To identify the Heckman selection model, we generate a variable which we argue is 1) relevant and 2) meets the exclusion restriction. Our generated variable (i.e., the instrument in the Heckman selection model) is a function of the changes in school sizes among the set of schools to which an applicant submits applications. More specifically, for each applicant, we examine the set of applications submitted in the most recent calendar year in which they apply to be a new teacher in WCPSS. For each school, we calculate the change in school size between the previous and current school year, measured in hundreds of students. [2] Then, to construct this variable, for each applicant, we assign the maximum value of the change in school size among the set of schools to which (s)he applied. This maximum value, which is a measure of the extent of growth in student numbers among the set of schools to which an applicant submits applications, is predictive of the likelihood that the applicant is hired in WCPSS. Schools that happen to experience larger population growth need to hire more teachers, so teachers who have a school with large student population growth in their set are more likely to be hired as new teachers in WCPSS. In other words, we argue that this measure is a relevant instrument, as evidenced by results in Table A1, which we discuss below.

With this instrument, we estimate the following Heckman selection model:

$$Y_{jk} = \phi_0' + \phi_1' \text{Score}_j + v_{jk}' \text{ (observed only if } hired_j = 1\text{)}, \tag{A2}$$

$$hired_j^* = \varphi_0 + \varphi_1 Z_{jk} + \varphi_2 \text{Score}_j + w_{jk}, \tag{A3}$$

$$hired_j = \begin{cases} 1, & hired_j^* > 0 \\ 0, & \text{otherwise} \end{cases} \tag{A4}$$

where $Z_{jk}$ is the excluded variable (i.e., the measure of the extent of school growth among schools applied), and $hired_j^*$ is the propensity for applicant $j$ to be hired. If $hired_j^*$ is greater than the threshold value (set at 0), then $hired_j = 1$ and we observe applicant $j$'s outcome $Y_{jk}$. We estimate this Heckman selection model using a two-step method (Heckman, 1979). $\phi_1'$ is the sample selection-corrected estimate of the relationship between the TeacherFit score and the outcome of interest.

Table A1 provides estimates from a Probit model that are analogous to first-stage estimates of the Heckman selection model. The outcome variable is being a newly hired teacher,

[2] WCPSS experienced large student population growth in this period, growing from 155,184 students in 2014-15 to 160,429 students in 2017-18.

and the predictors are the standardized TeacherFit sum score and our constructed variable (i.e., the measure of the extent of growth in student numbers among the set of schools an applicant submits applications). This model is executed at the applicant-level, and we report the average marginal effect from the Probit regression. Here, we see that applicants who have a higher measure of maximum school size growth among their set of schools are more likely to be hired as new teachers in WCPSS (F-stat. = 26.91; $p < 0.001$)—evidence that the instrument is relevant.

To meet the exclusion restriction, this measure needs to be related to the likelihood that the applicant is selected into the sample but otherwise be unrelated to the applicant's on-the-job performance. The exclusion restriction would be violated if higher quality applicants systematically have higher (or lower) values of this measure—the extent of student population growth among the schools to which they apply. If, for example, more motivated applicants (i.e., those who would have better performance or retention outcomes as teachers) systematically sought to apply to schools that have grown in school size while less motivated applicants did not, the exclusion restriction would be violated. That is, the extent of growth in school size among the schools to which individuals apply would be correlated with applicants' outcomes as teachers in this scenario. However, prior research by Cannata (2010) suggests that teacher applicants emphasize familiarity and location in their job search, avoiding schools with which they are unfamiliar or where they perceive they would feel uncomfortable. Cannata (2010) also finds that teacher applicants are less likely to research unfamiliar schools, often relying on assumptions based on information about student demographics and achievement. Such evidence suggests that this scenario—in which particularly motivated teacher applicants systematically submit applications to schools growing in size—is unlikely.

Another potential challenge to the validity of our instrument is the concern that teachers with higher values of the measure (i.e., the extent of student population growth among the schools in individuals' application sets) may be more likely to land teaching jobs in schools with lower or higher performance standards or mobility outcomes, as schools that experience growth in school size might be systematically different from those that do not grow. In this scenario, the instrument could be correlated with teachers' outcomes by way of the systematic differences in the characteristics of the schools in which the applicants find jobs. To shed light on whether such a scenario may be cause for concern, we use data from 2013-14 through 2017-18 to examine whether schools which experience larger prior changes in school size have, on average, higher or lower mean teacher outcomes. We fit the following simple regression:

$$Y_{st} = a + b * \text{Size}_{s,t-2,t-1} + e_{st}, \tag{A5}$$

where $Y_{st}$ is the school-level mean of teacher outcome $Y$ at school $s$ at time $t$; $\text{Size}_{s,t-2,t-1}$ is the change in student population at school $s$ between time *t-2* and *t-1*, and $e$ is the error term. Standard errors are clustered at the school-level.

We present the regression results in Table A2. We do not find a significant relationship between prior school size change and the mean teacher evaluation, retention, and value-added outcomes (Columns 1, 3-6). However, we do find evidence to suggest that schools with larger prior changes in school size, on average, have lower mean teacher absences (Column 2). These results perhaps cast some doubt on the validity of the instrument for the Heckman selection model in which teacher absences serve as the outcome of interest, but we do not find evidence to doubt the validity of the instrument for the selection models focused on the other main teacher outcomes.

If the exclusion restriction were indeed violated—for the reasons described above or any other—such that the instrument is related to applicants' on-the-job performance, the selection corrected estimates would also be biased. In other words, without a valid instrument, we would not be able to alleviate concerns about bias in the estimates due to sample selection. However, other studies (e.g., Goldhaber et al., 2017; Jacob et al., 2018) which generate sample selection corrected estimates in similar contexts also find little difference between sample selection corrected and uncorrected estimates, suggesting that selection bias may be fairly small. The lack of evidence of substantial selection bias in other similar studies, in conjunction with the similarity between corrected and uncorrected estimates in our study (Table 3 in main text), provide some assurance that bias from selection into the sample is likely to be small if present.

However, as discussed by Jacob et al. (2018), the direction, or sign, of the bias due to sample selection cannot be determined a priori, as one can easily imagine plausible narratives that would lead to bias in either direction. For example, applicants with low scores—who are hired despite their low scores—may be hired because they are strong in other areas that are positively related to teacher performance but are unobservable to us researchers. This would lead to positive selection of applicants in the lower end of the score distribution, thereby biasing our estimates towards zero. On the other hand, it is also possible that applicants who have high scores, but low commitment to teaching, choose to reject job offers in WCPSS in favor of outside (non-teaching) options. Assuming that commitment to teaching is predictive of teacher performance, those who end up accepting jobs in WCPSS from the upper end of the score distribution would be positively selected, thereby biasing our estimates upwards. Such plausible scenarios about selection into the sample make the direction of the bias unclear.

## References

Applitrack. (2015). Choosing the Best Applicant Screening Tool. Retrieved from http://www.frontlinek12.com/FrontlineK12/media/images/resources-pdfs/AppliTrack_TeacherFit_FitVsOthers.pdf

Applitrack (n.d.). *Example School District 123 – Employment Application.* Retrieved April 10, 2018, from https://www.applitrack.com/dex/onlineapp/_application.aspx?posJobCodes=117&posFirstChoice=AppliTrack%20Fit%20Trial&posSpecialty=

Cannata, M. (2010). Understanding the teacher job search process: Espoused preferences and preferences in use. *Teachers College Record, 112*(12), 2889-2934.

Cannata, M., & Engel, M. (2012). Does charter status determine preferences? Comparing the hiring preferences of charter and traditional public school principals. *Education Finance and Policy*, *7*(4), 455-488.

Cannata, M., Rubin, M., Goldring, E., Grissom, J. A., Neumerski, C. M., Drake, T. A., & Schuermann, P. (2017). Using teacher effectiveness data for information-rich hiring. *Educational Administration Quarterly*, *53*(2), 180–222.

Engel, M., Cannata, M., & Curran, F. C. (2018). Principal influence in teacher hiring: Documenting decentralization over time. *Journal of Educational Administration, 56*(3), 277-296.

Frontline Education (2022). Application Screening Assessments. Retrieved June 28, 2022, from https://www.frontlineeducation.com/solutions/recruiting-hiring/screening-assessments/

Frontline Technologies | K-12 Administrative Software Solutions. (2015). Retrieved September 27, 2015, from http://www.frontlinek12.com/Home.htmlGallup Online.

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

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153. https://doi.org/10.2307/1912352

Hinrichs, P. (2021). What kind of teachers are schools looking for? Evidence from a randomized field experiment. *Journal of Economic Behavior & Organization*, *186*, 395-411.

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

Johnson, S. M. (2020). *Where teachers thrive: Organizing schools for success*. Harvard Education Press.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

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

McKinney, J. C. (2017). *Who Gets the Job? Who Keeps the Job? An Analysis of the Relationship between Applicants' Performance on the AppliTrack Pre-Screening Tool and the Likelihood of their Hiring and Retention in One School District.* [Doctoral dissertation, University of Kansas]. KU ScholarWorks.

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

Oestreich II, T. A. (2016). *A Study Of The Effectiveness Of The Applitrack Teacherfit Hiring Selection.* [Doctoral dissertation, Indiana University]. IUScholarWorks.

Perrone, F., & Meyers, C. V. (2021). Teacher Hiring in the United States: A Review of the Empirical Research (2001-2020). (EdWorkingPaper: 21-459). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/58hw-zn20

Reese, S. (2018). *Applitrack Selection Assessments* [Webinar]. Frontline Technologies. https://www.youtube.com/watch?v=Tl7ngKG1xoI

Rose, D. S., English, A., & Finney, T. G. (2014). *Hire Better Teachers Now: Using the Science of Selection to Find the Best Teachers for Your School*. Harvard Education Press.

Simon, S. (2014, December 29). Teacher hopefuls go through big data wringer. *Politico*. https://www.politico.com

Young, I. P., & Delli, D. A. (2002). The validity of the teacher perceiver interview for predicting performance of classroom teachers. *Educational Administration Quarterly*, *38*(5), 586-612.

Table A1: First-stage Probit Estimates

| | Hired |
|---|---|
| Std. TeacherFit Sum Score | 0.018*** |
| | (0.004) |
| Measure of sch. size growth | 0.009*** |
| | (0.002) |
| | |
| N (applicants) | 10,714 |
| F (excluded variable) | 26.91 |

Notes: *** p<0.001. Estimates are average marginal effects from a Probit regression.

Table A2: Relationship between Prior School Size Changes and Mean Outcomes

| | Std. Evaluation Score | Absences | Within School Retention | Within District Retention | Math VA | ELA VA |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| School size change from | -0.005 | -0.555*** | 0.000 | -0.001 | 0.002 | 0.002 |
| t-2 to t-1 (in hundreds) | (0.023) | (0.133) | (0.005) | (0.002) | (0.007) | (0.004) |
| | | | | | | |
| N (school-years) | 838 | 838 | 838 | 838 | 708 | 707 |

Notes: *** p<0.001. Standard errors are clustered at the school level.