CAN YOU RECOGNIZE AN EFFECTIVE TEACHER WHEN YOU RECRUIT ONE?

Jonah E. Rockoff
(corresponding author)
Graduate School of Business
Columbia University
New York, NY 10027
jonah.rockoff@columbia.edu

Brian A. Jacob
Gerald R. Ford School
of Public Policy
University of Michigan
Ann Arbor, MI 48109
bajacob@umich.edu

Thomas J. Kane
Harvard Graduate School
of Education
Appian Way, Gutman 455
Cambridge, MA 02138
tom_kane@harvard.edu

Douglas O. Staiger
Department of Economics
Dartmouth College
Hanover, NH 03755
Douglas.O.Staiger@
dartmouth.edu

Abstract

Research on the relationship between teacher characteristics and teacher effectiveness has been underway for over a century, yet little progress has been made in linking teacher quality with factors observable at the time of hire. To extend this literature, we administered an in-depth survey to new math teachers in New York City and collected information on a number of nontraditional predictors of effectiveness, including teaching-specific content knowledge, cognitive ability, personality traits, feelings of self-efficacy, and scores on a commercially available teacher selection instrument. We find that only a few of these predictors have statistically significant relationships with student and teacher outcomes. However, the individual variables load onto two factors, which measure what one might describe as teachers' cognitive and noncognitive skills. We find that both factors have a moderately large and statistically significant relationship with student and teacher outcomes, particularly with student test scores.

> And this is our present purpose: to discover, so far as possible, what elements enter into the making of a capable teacher.
>
> —*J. L. Meriam, Teachers College Contributions to Education No. 1 (1906)*

1. INTRODUCTION

Research on the relationship between teacher characteristics and teacher effectiveness has been underway for over a century, yet little progress has been made in linking teacher quality with factors observable at the time of hire (see reviews by Hanushek 1986, 1997; and Greenwald, Hedges, and Laine 1996). Teaching experience is perhaps the only characteristic that has consistently been found related to teacher effectiveness, but a recruitment policy of hiring only veterans would be infeasible in most school districts. At the same time, the importance of recruiting high-quality teachers has been bolstered by recent work demonstrating substantial and persistent variation in achievement growth among students assigned to different teachers (e.g., Rockoff 2004; Rivkin, Hanushek, and Kain 2005; Kane, Rockoff, and Staiger 2008; Aaronson, Barrow, and Sander 2007). These findings have led to proposals that districts pay more attention to performance in the early part of teachers' careers as opposed to spending more resources on recruitment and hiring (Gordon, Kane, and Staiger 2006).

However, most research on teacher effectiveness has examined a relatively small set of teacher characteristics, such as graduate education and certification, which are collected by school administrators in order to satisfy legal requirements and set salaries. Like the well-known story of a man looking for his keys under a street light—not because he dropped them nearby, but because that is where he can see—researchers' lack of success in predicting new teacher performance may be driven by a narrow focus on commonly available data.

In the present study, we explore whether certain characteristics not typically collected by school districts can predict teacher effectiveness. To do so, we administered an in-depth survey of new elementary and middle school math teachers in New York City in the school year 2006–7. The survey assesses a host of teacher qualities at the time of hire, including general cognitive ability, content knowledge, personality traits (e.g., extraversion), and personal beliefs regarding self-efficacy. We match this survey data to administrative data on students and teachers in New York City, which allows us to explore how both traditional (e.g., certification type, teacher certification exam scores, selectivity of undergraduate institution) and nontraditional measures of teacher effectiveness predict five outcomes: the achievement of teachers' students on

standardized math tests, subjective teacher performance ratings, teacher absences, and teacher retention at both district and school levels. In addition to comparing the predictive power of our nontraditional measures with the several traditional measures, we also explore how well sets of variables can jointly predict teacher effectiveness.

We then investigate a commercial instrument—the Haberman Star Teacher Evaluation PreScreener (hereafter Haberman PreScreener)—whose purpose is to provide school officials with guidance on how effective a particular candidate is likely to be in an urban classroom (Haberman 1993, 1995). The Haberman PreScreener, developed in the 1980s, has been used by many urban school districts throughout the United States. We examine what teacher characteristics are associated with high scores on the Haberman PreScreener and then test whether performance on this instrument predicts a variety of teacher and student outcomes.

We find modest and marginally significant relationships between student achievement and several nontraditional predictors of teacher effectiveness, including performance on the Haberman selection instrument and a test of math knowledge for teaching. Interestingly, we do not find that respondents' levels of conscientiousness or extraversion (as measured on a standard personality inventory) are significantly related to student achievement, but they are strong predictors of subjective evaluations made of respondents. This finding is of interest given a large literature on the impacts of worker personality on job performance, which often uses subjective evaluations by supervisors as the performance metric.

While no single metric we examine has the ability to reliably identify very large differences in teacher effectiveness, we document how these metrics can be used to create composite measures of cognitive and noncognitive skills, both of which have statistically significant relationships with student achievement. Together these factors have modest but economically meaningful power for screening effective teachers at the time of hire. Our estimates suggest that students assigned to a teacher who is 1 standard deviation higher on either the cognitive or noncognitive factor have achievement that is .025 student-level standard deviations higher. In comparison, in prior work we have found that a value-added measure based on a teacher's performance over the first two years of his or her career has an effect size of .072 in predictions of subsequent student math achievement. These results suggest that schools and school districts wishing to increase the effectiveness of their teacher workforce may benefit from gathering a broad set of information on new candidates but that data on job performance may still be a more powerful tool for improving teacher selection than data available at the recruitment stage.

The article proceeds as follows. In section 2 we describe the data used in our analysis and provide descriptive statistics. In section 3, we present our methodology. Section 4 presents our findings, and section 5 offers some conclusions.

2. DATA

The main focus of our analysis is an online survey of teachers who began their careers in New York City public schools in the school year 2006–7. In this section, we describe survey implementation, present brief descriptions of survey elements (focusing on nontraditional items), and examine the additional administrative data used in our analysis. We then discuss sample selection and present descriptive statistics of our analysis sample.

Teacher Survey

With the assistance of school district officials, we identified all individuals with no prior experience who were listed as teaching mathematics to students in grades 4–8 in the 2006–7 academic year ($N = 602$). We limited our sample to math teachers in these grades so that we would be able to calculate a value-added measure of teacher effectiveness using at least one prior test score as a control (testing begins in third grade in New York City). We limited our sample to new teachers because of our interest in predicting effectiveness during initial hiring. We focused solely on math teachers (and not reading or science teachers, for example) due to budget constraints. It is worth emphasizing that we surveyed the entire population of new math teachers in these grades in this year.

Ideally we would have administered the survey to these teachers prior to the start of the school year. However, data linking students and teachers in New York do not become available until well past the start of the school year. In addition, some of the survey elements required us to navigate legal copyright issues, and this caused some delay. In the end, survey invitations went out on 3 April 2007, and teachers were given until the end of June to complete the survey.[1] The timing of the survey has implications for the interpretation of our results, and we discuss this further below.

The survey was fairly extensive, with seven parts and over two hundred items. In order to compensate teachers for this substantial amount of time, we offered a $75 payment for successful completion of the survey. Several reminders were sent to nonrespondents and noncompleters between the start

1. In order to protect the confidentiality of the data, communication with teachers was done via the school district's human resources department. Survey invitations contained a unique link, based on a scrambled teacher identification number, so that survey responses could be merged with other sources of data.

and end of the survey period. Of the 602 teachers invited to complete the survey, 418 (69.4 percent) began the survey, and 333 (55.3 percent) completed it entirely.[2]

The goal of this survey was to capture a set of information that has not been widely studied in the literature on teacher effectiveness but has been linked to teacher productivity or productivity in other occupations by prior research. Here we briefly review the major survey components included in our analysis.[3]

A Teacher's Cognitive Ability and Academic Success

We collected a number of common measures of a teacher's cognitive ability and academic success, including undergraduate major, graduate education, selectivity of undergraduate institution, and college entrance scores (i.e., ACT or SAT).[4] We also asked respondents about success on the Liberal Arts and Science Test (LAST), which is the primary teacher certification exam used in New York State.[5] There is some prior evidence that these measures are correlated with student performance, although in no case is the evidence dispositive. For example, some researchers have found that teachers with stronger academic backgrounds produce larger performance gains for their children (see, e.g., Clotfelter, Ladd, and Vigdor 2006, 2007), but others do not find this relationship (see, e.g., Harris and Sass 2006 on graduate course work and Kane, Rockoff, and Staiger 2008 on college selectivity). Similarly, Clotfelter, Ladd, and Vigdor (2006, 2007) and Goldhaber (2007) find a link between teacher certification scores and student performance and teacher effectiveness, while Harris and Sass (2006) do not.

One problem with interpreting the relation between successful teaching and college entrance exam scores is that performance on standardized achievement tests is determined by a host of different factors: access to educational

2. Respondents include all teachers who began the survey, including fifteen teachers who began the survey but did not complete any of the main sections. Placing these fifteen teachers in the nonrespondent category does not noticeably affect our comparisons of respondents and nonrespondents (table 1).
3. Note that we do not review the extensive literature on more traditional predictors of teacher effectiveness, which focuses on characteristics such as experience or certification type. For a review of this literature, see Jacob (2007).
4. We asked respondents for their undergraduate institution, and we merged this information with the Barron's Selectivity Index (a 1–9 scale, 1 being the best) from 1982. We thank Caroline Hoxby for sharing these data with us. For a few colleges where the Barron's rating was missing, we used Barron's ratings from 1984. There is no comprehensive selectivity index by major within institution, so the index we use here is based on all undergraduates at each institution.
5. In addition to the LAST exam, teachers may also be required to pass the Assessment of Teaching Skills (ATS-W), and a Content Specialty Test (CST) may also be required, depending on subject area and certification type. For example, the ATS-W is not required of alternatively certified teachers (e.g., TFA and teaching fellows). We do not present results on the predictive power of these exam scores, but these results are available upon request. In preliminary analyses, we found that exam scores had no significant power to predict student achievement, and the point estimates are very small and, in some cases, of the wrong sign.

resources in childhood, parental investment in education, personal motivation and willingness to study hard, raw intelligence, etc. In order to separate at least one of these proximate causes, the survey includes a direct test of cognitive ability, Raven's Progressive Matrices Standard Version (Raven's test), an intelligence test that requires no linguistic or mathematics skills.[6]

Content Knowledge

A number of studies examine the relationship between content knowledge and teacher effectiveness, particularly in mathematics (e.g., Goldhaber and Brewer 1997; Aaronson, Barrow, and Sander 2007). Although the evidence on this issue is mixed, these studies use proxies for content knowledge, such as the number of courses taken in a subject or college major. Some math educators and researchers argue that it is not simply mathematical knowledge per se, but the ability to express mathematical concepts in the context of classroom teaching that is critical. Mathematical knowledge for teaching involves the ability to explain difficult mathematical concepts in multiple ways and to describe the intuition behind mathematical reasoning instead of focusing exclusively on algorithms and procedures (Shulman 1986, 1987; Wilson, Shulman, and Richert 1987).

Motivated by this work, we measure content knowledge using an instrument developed by researchers at the University of Michigan designed to assess this specific type of knowledge (Hill, Rowan, and Ball 2005). Hill, Rowan, and Ball (2005) find that this measure is positively correlated with student achievement gains in first and third grades and that it is a stronger predictor of student learning than other measures of teachers' mathematical preparation.

Personality Traits

There is a long history of studying teacher personality characteristics in the education literature (see a review by Getzels and Jackson 1963). While many studies examine the relationship between teacher personality and student performance, the vast majority have utilized the Minnesota Multiphasic Personality Inventory (MMPI), an instrument that has been criticized on various grounds. In other work, however, psychologists have used a more commonly accepted framework for measuring personality traits known as the five-factor model (or the "Big Five") to successfully predict job performance across a wide variety of occupations. The Big Five personality traits are agreeableness,

6. The test relies on the participant's ability to recognize and decode patterns of symbols presented in a matrix. Each set of items becomes progressively more difficult, requiring greater cognitive capacity to encode and analyze (Raven 2000). Though it has been found to have a high correlation with other major tests of intelligence (Raven and Summers 1986), it is considered to be one of the best measures of general cognitive ability due to its nonverbal nature. The split-half reliabilities for this test are also high, with a coefficient of .86 (Raven, Court, and Raven 2000).

conscientiousness, emotional stability, extraversion, and openness to experience. While we are not aware of any work linking elements of the Big Five to teacher effectiveness in raising student achievement, Barrick and Mount (1991) find that conscientiousness is positively related to job performance across all occupational categories. They also document a link between extraversion and job performance in occupations requiring social interaction.[7]

Instruments used to measure the Big Five vary in length and complexity. We employ the Big Five Inventory (BFI), developed by John, Donahue, and Kentle (1991), which consists of forty-four items: ten for openness to new experience, nine each for agreeableness and conscientiousness, and eight each for emotional stability and extraversion. Each item asks respondents for their level of agreement (on a scale of 1 to 5) with a statement about themselves, and about half the items are reverse scored. For example, agreement with the statements "I am someone who is talkative" and "I am someone who is reserved" are both used to measure extraversion, but the latter is reverse scored. Each respondent receives a score from 1 to 5 on each of the five dimensions of personality.

Teacher Beliefs and Values

A number of researchers have examined variation in teacher self-efficacy and its correlation with student and school outcomes (e.g., Gibson and Dembo 1984; Dembo and Gibson 1985; Woolfolk and Hoy 1990; Raudenbush, Rowan, and Cheong 1992; Hoy and Woolfolk 1993). This body of work generally finds a positive relationship between self-efficacy and outcomes such as supervisor ratings, even after controlling for some potentially confounding covariates. However, there is little work examining the relationship between self-efficacy and student learning.[8]

Following the prior work on teachers' self-efficacy, we measure self-efficacy in two ways: personal efficacy (i.e., belief in one's own ability to affect student learning) and general efficacy (i.e., belief in the ability of teachers in general to affect student learning). We use a ten-item instrument developed by Hoy and Woolfolk (1993), adapted from earlier work by Gibson and Dembo (1984). A simple factor analysis of teachers' responses finds two factors, with the general and personal efficacy items grouped as expected.

7. Similar results are echoed in a review by Goodstein and Lanyon (1999).
8. One exception is an often overlooked result in a well-cited study on teacher quality by Armor et al. (1976). In addition to being one of the first studies of teacher value added and its correlation with principal evaluations, this article also finds a significant positive relationship between teachers' sense of self-efficacy and student achievement growth.

Teacher Selection Instruments

One policy goal of research on predictors of teacher effectiveness is to develop tools that district and school administrators could use to identify the promising teacher candidates. However, there are already two commercially available and widely used instruments designed to measure beliefs and values indicative of future success in the classroom: the Haberman PreScreener and the Gallup TeacherInsight Assessment (Gallup TIA). Both instruments use a short survey consisting mostly of multiple-choice items to assess a number of teacher attributes that were exhibited by highly effective teachers. These instruments have been used by many large urban school districts throughout the United States, including Atlanta, Buffalo, Cleveland, Dallas, Denver, Long Beach, Los Angeles, Minneapolis, Nashville, Philadelphia, Pomona, San Francisco, San Diego, Tampa, and Washington, DC.

While use of commercial selection instruments has grown considerably, there is little systematic evidence on the power of these instruments for predicting teacher effectiveness. In this article, we analyze the Haberman PreScreener. This instrument uses fifty multiple-choice items to assess ten different attributes: persistence, organization and planning, beliefs about the value of students learning, approach to students, approach to at-risk students, ability to connect theory to practice, ability to survive in a bureaucracy, fallibility, explanation of students' success, and explanation of teacher success.[9] The Haberman instrument was included as a part of our survey and was scored for us by the Haberman Foundation.

Administrative Data

Administrative data from the New York City Department of Education (DOE) payroll system provides us with information on all full-time teachers in the DOE in September, November, and May of each school year since 1999–2000. These data include each teacher's gender and ethnicity, certification route/program (i.e., whether a teacher was traditionally certified or entered via an alternative certification program such as Teach for America [TFA] or the New York City teaching fellows), teaching experience (as proxied by their position on a salary schedule), number of absences, and whether they have left the DOE or switched schools.

We measure student achievement using data on standardized test scores in math for students in grades 4–8. These data follow students over time and provide links to their math teachers. The student data we possess also

9. In addition to the items described above, we also asked about several other characteristics (e.g., prior occupations, prior experiences working with children). In preliminary analyses not reported here (but available upon request), we found no systematic and/or significant relationship between these measures and our outcomes.

include information on demographics, receipt of free and reduced price lunch, and status for special education and English language learner services. A full description of the data can be found in Kane, Rockoff, and Staiger (2008).

One of the outcomes we examine is a subjective evaluation of teacher effectiveness by a mentor who meets with the teacher weekly and makes classroom observations. A small but growing literature demonstrates a significant relationship between objective measures of teacher performance and subjective evaluations of teacher quality made during a teacher's career (e.g., Murnane 1975; Armor et al. 1976; Harris and Sass 2008; and Jacob and Lefgren 2008). We obtain these mentor evaluations from a centrally administered program to assist new teachers (see Rockoff 2008).

Mentors are each assigned a group of roughly 15–20 teachers, usually spread across a number of different schools. In addition to working with teachers, mentors submit monthly summative evaluations of teachers' skills on a five-point scale ranging from "beginning" to "innovating." In practice, almost all teachers are rated "beginning" at the start of the school year, and some teachers are missing ratings for a subset of months. In order to have meaningful variation in evaluations, we concentrate on evaluations submitted toward the end of the year.[10]

In order to control for observable school characteristics in some of our analyses, we collected school-level information from the National Center for Education Statistics (NCES) Common Core of Data (CCD). This includes school-level data on student ethnicity, gender, and student eligibility for free lunch, as well as the school's pupil-teacher ratio and grade composition. In order to better control for differences across schools that are unobservable in the CCD data but are related to local neighborhood characteristics, we identified the zip code of each school in our sample, which allows us to include school zip code fixed effects.

Descriptive Statistics

Table 1 provides summary statistics broken down into three groups: new math teachers in grades 4–8 who responded to our survey (respondents, $N = 418$), new math teachers in grades 4–8 who did not respond to our survey (nonrespondents, $N = 184$), and other new teachers hired in 2006–7 (in other grades and/or subjects) that were not invited to participate in the survey ($N = 4,275$). The third column provides p-values on tests of whether there is a statistically

10. To avoid bias due to either the timing of evaluations or the leniency of mentors, we subtract the average rating given by each mentor in each month from an individual teacher's rating (i.e., we normalize ratings by mentor-month cell). We then average ratings given in the months of April, May, and June. For the teachers who were not rated in those months (less than 2 percent of teachers with any recorded evaluations), we use ratings averaged over January, February, and March.

Table 1. Comparison of Teachers by Survey Invitation and Response

	Respondents	Nonrespondents	Test of Equality by Response (*p*-value)	Not Invited to Survey
Number of teachers	418	184		4,275
Outcomes				
Teacher absences	5.70	5.76	0.87	6.40
Mentor rating overall	0.04	−0.05	0.07	−0.01
Teacher returned to NYC	91.9%	93.5%	0.49	92.6%
Teacher returned to school	83.0%	85.3%	0.48	84.5%
Teacher characteristics				
Female	77.8%	66.3%	0.00	75.5%
Black	13.9%	17.4%	0.27	13.1%
Hispanic	8.6%	9.2%	0.80	11.6%
Asian	9.8%	9.2%	0.83	6.3%
Age	27.74	27.01	0.18	28.62
Traditionally certified	48.8%	46.2%	0.56	51.5%
Teaching fellow	29.2%	25.0%	0.29	31.3%
Teach for America member	14.8%	22.3%	0.03	8.3%
Master's degree	31.3%	25.5%	0.15	35.8%
School characteristics				
Percent black	34.1%	36.9%	0.25	35.6%
Percent Hispanic	47.9%	47.7%	0.95	45.0%
Percent Asian	9.7%	7.8%	0.17	9.0%
Pupil-teacher ratio	14.34	14.28	0.75	14.51
Percent free lunch	75.2%	75.3%	0.97	70.2%

Notes: Shown are the average values of each variable, broken down by whether teachers were invited to take the survey and whether they responded to the invitation. School characteristics are taken from the Common Core of Data. *P*-values are taken from a test of the significance of an indicator for survey response in a regression that includes only those individuals who were invited to take the survey.

significant difference in the mean of a characteristic between respondents and non-respondents.

Of the eighteen teacher and school characteristics listed in the table, there are two on which the respondents and nonrespondents are significantly different at the 5 percent level or lower. Relative to nonrespondents, respondents were more likely to be female (78 percent vs. 66 percent) and were less likely to come from the TFA program (15 percent vs. 22 percent). Though the *p*-value is slightly above 0.05, it is also noteworthy that survey respondents were given higher subjective evaluations by their mentors (0.04 vs. −0.05). Given that the individuals who were not asked to take the survey (column 4) were teaching in different grades and/or subjects than our survey sample, there is no reason

Table 2. Summary Statistics on Survey Responses

	Observations	Mean	Standard Deviation
Academic background			
Math/science major	403	20.6%	
Education major	403	14.6%	
Has a graduate degree	402	32.1%	
SAT verbal score	270	606.1	94.5
SAT math score	271	613.0	90.9
Barron's rank of college (1–9 scale, 1 is best)	248	5.6	1.9
Passed the LAST certification exam on first try	370	92.2%	
Cognitive ability (percentile)	333	53.4	25.9
Math knowledge for teaching (percent correct)	337	0.57	0.20
Personality			
Extraversion	396	3.60	0.66
Agreeableness	396	4.11	0.45
Conscientiousness	396	4.04	0.52
Emotional stability	396	4.44	0.64
Open to new experiences	396	3.85	0.53
Self-efficacy			
Personal efficacy	387	3.81	0.63
General efficacy	387	3.19	0.79
Haberman PreScreener performance			
Haberman top group	338	21.3%	
Haberman total correct	338	31.86	4.81

to believe that they should be identical to those teachers in columns 1 and 2, although in practice they do appear fairly similar.[11]

Table 2 presents summary statistics on variables from our survey, grouped by broad themes. The number of non-missing observations varies across survey items due to varying completion rates by respondents and the position of the item in the survey. The academic backgrounds of survey respondents are quite varied. Approximately one in five survey respondents majored in either math or science, and about one in six majored in education. The fairly high averages for SAT scores may reflect the percentage of teaching fellows and TFA corps members in our sample, and perhaps nonrandom selection in teachers' willingness to report their scores. Nearly all the respondents (92.2 percent) claimed to have passed the LAST exam on their first attempt.

The average score on the test of cognitive ability fell at the 53rd percentile relative to national norms. The standard deviation was 26 percentile points,

11. We have also compared the characteristics of teachers who completed the survey with those that began but did not complete (results available upon request). Relative to individuals who completed the entire survey, individuals that started but did not complete the survey were more likely to be nonwhite and less likely to come from the TFA program.

indicating a substantial amount of heterogeneity in cognitive ability in our sample. Indeed, the scores for survey respondents matched the national norms to within one point at the 25th, 50th, 75th, 90th, and 95th percentiles. They outperformed the national distribution at the 5th and 10th percentiles, but given that all these teachers must have a college degree, this is not terribly surprising.

The portion of correct answers on the test of math knowledge for teaching was 0.57 on average, with a standard deviation of 0.20. The 10th and 90th percentiles of respondents correctly answered 33 and 83 percent, respectively.[12] Scores on the math knowledge for teaching exam were positively correlated with self-reported math SAT ($r = 0.46$), verbal SAT ($r = 0.38$), cognitive ability ($r = 0.49$), and the (inverse of) Barron's selectivity rating of undergraduate institution ($r = 0.34$).

In table 2 we report the raw scores (on a scale of 1–5) for all five dimensions of personality from the BFI, though in our analysis below we restrict our attention to conscientiousness and extraversion. The summary statistics for these measures are difficult to interpret (e.g., a score of 3.8 on the agreeableness measure has no natural meaning), and, to the best of our knowledge, there are no national norms for these traits that use a five-point scale like we do. Therefore, rather than ask whether survey respondents score higher or lower than a national sample on a particular trait, we examine whether the ratio of a particular trait to the other traits among our survey respondents is greater or less than ratios for a national sample. Using this (admittedly informal) method, we find that our survey respondents have relatively higher scores on emotional stability, lower scores on extraversion, and similar scores on conscientiousness, agreeableness, and openness to new experiences.[13] However, there are no striking differences between the two samples' scores.[14]

Finding a benchmark for the self-efficacy scores is also difficult, so we compare our survey respondents' average scores (3.8 for personal efficacy and 3.2 for general efficacy) to samples in the prior literature. Our respondents' scores are lower than teachers surveyed by Woolfolk and Hoy (1990) and Hoy and Woolfolk (1993), where samples averaged, respectively, 4.2 and 4.7 for general efficacy and 3.6 and 3.8 for personal efficacy, and unlike in their sample, our survey respondents score higher on personal efficacy. However, the variation in scores within all three groups is of similar magnitude. The

12. In addition to the portion answered correctly, we estimated scaled scores for this test using item response theory. The results of our analysis are quite similar using the scaled scores or the portion correct, and thus, for greater transparency, we report results for the portion correct.
13. The mean scores for the nationally representative sample on the 1–4 scale were 3.48 for agreeableness, 3.42 for conscientiousness, 3.20 for extraversion, 2.76 for emotional stability, and 3.02 for openness to new experiences.
14. Note that we do not use the ratio of traits in the analysis below. The ratios are merely presented here as a way to compare our sample with a national sample.

correlation between personal and general efficacy in our sample is 0.15, which is identical to the sample in Hoy and Woolfolk (1993) and similar to the correlation of 0.07 found for the sample in Woolfolk and Hoy (1990).

Among teachers who completed the Haberman PreScreener, just over 20 percent fell into the top group of candidates according to the recommended classification system. The median total number of items answered correctly (out of fifty) was thirty-two, which is identical to the median reported by Haberman for individuals in other districts that have completed the instrument. The standard deviation of the number correct was 5.

Our Analysis Sample

While our analysis focuses on the 418 teachers who responded to our survey, we include other teachers in our analysis in order to help identify coefficients on variables other than those from our survey (e.g., student and school characteristics). Specifically, when examining teacher outcomes (e.g., subjective evaluations, absences, and retention) we include data on 1,190 other new, inexperienced teachers working in the same school as at least one survey respondent, 84 of whom were asked to take the survey but did not respond. For each of the outcomes, our sample naturally includes only those teachers with valid outcome data. We have attrition data for all 1,608 teachers in our sample and absence data for all but four of these teachers. We have mentor ratings for 1,117 teachers (69 percent of our sample). The missing mentor data stem from the fact that roughly one-quarter of the schools in the district received an exemption from the centralized mentoring program due to their status as Empowerment Schools (a program that gave principals more programmatic autonomy).[15]

For our analysis of student achievement, we use a slightly different sample. Specifically, we include all students and teachers in the value-added grades (grades 4–8) during the school year 2006–7 working in schools with at least one survey respondent. We include teachers (and their students) who have been in the district for any number of years, and not just new teachers. We include these additional classrooms in order to gain better estimates of the coefficients on important control variables, such as prior student achievement,

15. Over 60 percent of the missing evaluations are due to teachers working in Empowerment Schools. Of the remaining teachers, 86 percent are merged with data from the mentoring program, which is in line with earlier program years (see Rockoff 2008) and is likely due to administrative errors and late hiring. The fraction of teachers with mentor evaluations is somewhat higher among teachers who responded to our survey (75 percent) or who were asked to take our survey but did not respond (70 percent) than among those who were not asked (67 percent). The fraction of teachers with mentor evaluations among teachers not in empowerment schools is also higher among teachers who responded to our survey (91 percent) or who were asked to take our survey but did not respond (88 percent) than among those who were not asked (84 percent).

participation in English language learner and special education programs, etc. In addition, we implement the sample restrictions used in Kane, Rockoff, and Staiger (2008), which eliminates classes with high fractions of special education students or where the teacher-student match may be incorrect.[16]

For both student and teacher outcomes, we have run additional analyses restricting our sample to (students of) teachers who responded to our survey. The coefficients on our predictor variables were similar to those presented below, and when they differed they tended to be larger, though, not surprisingly, less precisely estimated. Thus we prefer the specifications that use a broader sample of teachers and students.

3. EMPIRICAL STRATEGY

Our primary goal is to determine which, if any, measurable teacher characteristics predict various teacher and student outcomes. When we consider teacher-level outcomes (e.g., number of teacher absences in a given year, mentor's rating of the teacher), we will estimate a regression like the one shown by equation 1, where Y_{jk} is the outcome for teacher j in school k, P_j is a predictor of teacher effectiveness, X_j (SC_{jk}) are other teacher (school) characteristics that are included as control variables in certain specifications, and ε_{jk} is an idiosyncratic error term

$$Y_{jk} = \alpha + \delta P_j + \beta X_j + \gamma SC_{jk} + \varepsilon_{jk}. \quad (1)$$

More specifically, we include (1) controls for the characteristics of schools from the CCD including percent Asian, percent black, percent Hispanic, percent free lunch, and the pupil-teacher ratio, as well as indicators for the school level (as defined by the NCES: primary, middle, high, or other) and indicators for the highest grade level served by the school; (2) school zip code fixed effects, and (3) grade-level indicators for survey respondents.

When examining student achievement data, we estimate a similar specification (shown in equation 2) where A_{ijk} is the achievement level of student i, assigned to teacher j in school k, and S_i represents a set of controls for student characteristics, including prior achievement:

$$A_{ijk} = \alpha + \delta P_j + \beta X_j + \gamma SC_k + \lambda S_i + \varepsilon_{ijk}. \quad (2)$$

Following the approach described above, we include students taught by teachers who did not respond to the survey but worked in the same school as at

16. In total, we are unable to examine math value added for 39 of our 418 survey respondents: 7 were not linked to students in our testing data, 2 taught in schools for which we could not match at least 75 percent of students to teachers, and 30 taught in classrooms where more than 25 percent of the students were classified as receiving special education services.

least one survey respondent. To account for differences across classrooms, we include controls for individual students' prior student test scores (specifically, cubic polynomials in both prior math and reading scores, interacted with grade level) and student demographics (gender, ethnicity, participation in free lunch, special education, and English language learner programs, and the number of absences and suspensions in the prior school year). Instead of the CCD variables described above, we include classroom and school averages of these student characteristics, as well as teaching experience indicators, school zip code fixed effects, and grade-level fixed effects interacted with the lowest grade served by the school.[17] We regard this specification as generating valid estimates of the relationship between survey variables and teacher effectiveness. While we recognize that the inclusion of school fixed effects would be a more robust methodology, only 24 percent of the schools that had any survey respondents had more than one, making within-school identification impracticable.

We examine five dependent variables in our analysis: student achievement in math, subjective evaluations of teachers, teacher absences, whether a teacher returns to the DOE the following year, and whether a teacher returns to the same school the following year. Both test scores and subjective evaluations have been normalized to have a standard deviation of 1 so that coefficients can be readily interpreted.

When examining both student and teacher outcomes, we set predictor variables to zero for teachers with missing data and include a set of indicators for whether a student's teacher was not asked to take the survey, was asked but did not respond, or responded to the survey but did not complete the particular item being tested. Thus, while our sample size does not vary across the specifications, the true number of teachers with identifying variation fluctuates slightly depending on the number of teachers completing different portions of our survey. Again, data for these teachers/students are included in the regression in order to obtain better estimates of the coefficients on our control variables, and our results, while more precise with these larger samples, are not driven by their inclusion.

Before presenting our results, it is worth considering several issues with regard to how our estimates should be interpreted. First, even with our in-depth survey, we measure a limited set of teacher characteristics, and thus our models will miss many characteristics that might influence student learning (e.g., a teacher's empathy, toughness, love for children, personal charisma, connections to others with teaching experience, etc.). Hence one might be concerned that our analysis could suffer from a standard omitted variable bias. Suppose,

17. Motivation for these interactions comes from Rockoff and Lockwood (2010), who provide evidence that adolescent students in middle schools significantly underperform their peers in K–8 schools.

for example, that extraversion and empathy are positively correlated and both positively affect student achievement. In this case, the exclusion of empathy from our estimates may lead us to overstate the effect of extraversion on student performance. Because one could construct equally plausible examples in which predicted bias goes in the opposite direction, the direction of bias arising from this type of general specification error is ambiguous.

While this is a potential concern, recall that a key objective of our exercise is the identification of potentially effective measures for the purpose of hiring. In this respect, we are concerned entirely with predicting effectiveness, in which case a reliable correlation may still be useful for teacher hiring. If extraversion and empathy were strongly correlated in a pool of applicants, for example, one could improve student outcomes by hiring those with high levels of extraversion even if empathy were the factor that influenced student learning. One might be able to improve student outcomes even more if one knew the importance of empathy and could measure it, but this does not diminish the value of knowing the bivariate correlation between extraversion and student performance.[18]

A second and more serious concern stems from the fact that our analysis includes only those teachers who were hired to teach in the DOE, and not the full set of individuals who applied for teaching positions. To the extent that school and district officials are purposefully selecting teachers and can select the most effective candidates, the hiring process itself may introduce selection bias. For example, suppose that teacher conscientiousness were positively associated with student performance. In this case, one would expect schools to hire candidates with greater levels of conscientiousness, on average. However, if school officials hire a candidate with a low degree of conscientiousness, it is likely that this individual is particularly strong in some other way. Since we cannot observe and control for all other potential factors used in hiring that might influence student outcomes, this type of selective hiring on the part of school administrators will bias our results toward zero. However, this type of bias occurs only if the school district had access to better information than is observed in our data when they selected teachers. Although school district officials may have had access to additional information (e.g., from face-to-face interviews with teachers), they are unlikely to have had access to many of the measures we analyze.

A third concern stems from the timing of our survey. As noted earlier, a variety of logistical problems delayed the administration of the survey until April 2007. One might be concerned that some of our estimates reflect reverse

18. In addition, if one knew the true "structural" relationship between teacher characteristics and effectiveness, one might utilize professional development to enhance those characteristics that lead to effectiveness.

causality (i.e., a teacher's success or lack thereof during the school year might have influenced his or her survey responses, rather than the survey responses predicting relative success). This is not a concern for the background variables (e.g., type of certification, college attended) and is unlikely to be a large concern for predictors such as the personality measures that purportedly reflect more permanent individual traits. On the other hand, reverse causality is a particular concern with regard to the teaching efficacy measures. To the extent that the experience of teaching (and the successes or failures that come with it) influences how individuals respond to the Haberman instrument, one should be cautious about interpreting the coefficients on this measure as well.

4. FINDINGS

Table 3 shows how well our "traditional credentials" predict each of our five outcomes measures. Within each column, lines separate coefficient estimates from regressions in which we include a single predictor or a group of related predictors. Few of the measures reach traditional levels of statistical significance, which is not surprising given our small sample size (recall that we have at most 418 survey respondents in each regression). However, in a number of cases, the magnitudes of the coefficient estimates are economically important, and the *p*-values are below 0.2. For this reason we highlight several findings that are not significant by conventional standards, taking care in each case to clearly indicate the *p*-value.

Consistent with many other researchers, we find no significant relationship between graduate education and teacher effectiveness in raising student math scores (column 1); indeed, the point estimate is negative. We do not find that respondents who passed the main state certification basic skills exam—the LAST—on the first attempt are significantly more effective, but very few survey respondents (8 percent) reported failing this exam, and the coefficient is imprecisely estimated.[19] When comparing alternatively certified teachers with the traditionally certified among the survey respondents, we find that teaching fellows are less effective (−0.035 standard deviations, *p*-value = 0.16) and TFA corps members are more effective (0.04 standard deviations, *p*-value = 0.15).[20]

19. We also tested the predictive power of respondents' self-reported certification test scores, but in no case did these approach statistical significance (results available upon request).

20. While the result on TFA is consistent with other findings (Decker, Mayer, and Glazerman 2004; Boyd et al. 2006; Kane, Rockoff, and Staiger 2008), the negative finding for teaching fellows contrasts with earlier work (Boyd et al. 2006; Kane, Rockoff, and Staiger 2008). Nonrandom selection of survey respondents does not drive this result, as the coefficient does not change when we use identifying variation on all teachers who were asked to take the survey, as opposed to only survey respondents. However, the negative finding on teaching fellows does disappear when we use identifying variation in the certification pathway of all teachers—i.e., including teachers (both fellows and nonfellows) hired in earlier years. Thus it appears to be the case that either this particular

Table 3. Traditional Predictors of Teacher and Student Outcomes

	Math Achievement	Subjective Evaluation	Teacher Absences	Returned to NYC	Returned to School | NYC
Has a graduate degree	−0.025	0.112	−0.207	−0.056	−0.011
	(0.023)	(0.152)	(0.423)	(0.036)	(0.039)
	[0.266]	[0.463]	[0.624]	[0.125]	[0.770]
Passed LAST certification exam on first attempt (1=yes)	0.038	0.039	0.264	−0.069	−0.011
	(0.036)	(0.202)	(0.667)	(0.042)	(0.064)
	[0.289]	[0.847]	[0.692]	[0.106]	[0.862]
Teaching fellow (relative to traditionally certified)	−0.035	−0.172	0.974	0.118	−0.014
	(0.025)	(0.151)	(0.517)*	(0.033)**	(0.043)
	[0.162]	[0.257]	[0.060]	[0.000]	[0.746]
TFA corps member (relative to traditionally certified)	0.043	−0.016	−0.509	0.130	0.114
	(0.030)	(0.145)	(0.443)	(0.038)**	(0.040)**
	[0.148]	[0.915]	[0.251]	[0.001]	[0.004]
Math or science major (relative to those other than math, science, or education)	0.038	−0.088	−0.690	−0.064	−0.024
	(0.030)	(0.209)	(0.556)	(0.051)	(0.051)
	[0.208]	[0.676]	[0.215]	[0.211]	[0.640]
Education major (relative to those other than math, science, or education)	−0.020	−0.152	−0.124	−0.092	0.051
	(0.032)	(0.154)	(0.492)	(0.043)**	(0.041)
	[0.541]	[0.324]	[0.801]	[0.034]	[0.218]
Self-reported SAT math score (s.d. = 1)	0.012	0.024	−0.055	0.003	0.009
	(0.014)	(0.081)	(0.216)	(0.021)	(0.016)
	[0.397]	[0.763]	[0.798]	[0.870]	[0.572]
Self-reported SAT verbal score (s.d. = 1)	−0.008	0.074	0.087	0.020	0.012
	(0.014)	(0.089)	(0.231)	(0.022)	(0.023)
	[0.592]	[0.407]	[0.707]	[0.343]	[0.615]
Barron's rank of college (s.d. = 1)	0.013	−0.185	0.027	0.026	−0.007
	(0.011)	(0.091)**	(0.225)	(0.019)	(0.023)
	[0.252]	[0.043]	[0.905]	[0.158]	[0.043]
Control for student/school characteristics and zip code FE	√	√	√	√	√
Observations	82,977	1,117	1,604	1,608	1,495

Notes: Each set of coefficients (separated by solid lines) represents different regressions. See text for a full listing of the student and school characteristics used as control variables. Standard errors (in parentheses) are clustered by school; *p*-values for each coefficient are shown in brackets. *significant at 10%; **significant at 5%; √ denotes controls included in the regression.

Student test score growth was greater on average with respondents who majored in math or science (0.04 standard deviations, *p*-value = 0.2) and

group of teaching fellows is relatively less effective than earlier cohorts or the gains to experience for teaching fellows are greater than for other teachers. Although we cannot distinguish these two explanations without additional data, Kane, Rockoff, and Staiger (2008) present some evidence in support of the latter hypothesis.

slightly lower with respondents who majored in education (−0.02, p-value = 0.54); a test of the equality of these coefficients can be rejected with a p-value of 0.13. Respondents' self-reported SAT math score and selectivity of respondents' undergraduate institutions, as measured by the Barron's scale, have positive coefficients that do not reach conventional significance levels.

Turning to the teacher-level outcomes in table 3, the only traditional credential that is related to subjective evaluations is college selectivity. The coefficient estimate implies that respondents who attended colleges with a ranking 1 standard deviation above average received evaluations that were 0.2 standard deviations *lower* than those given to teachers with average college selectivity.[21] We find no statistically significant difference in the average evaluation given to respondents who were alternatively certified versus those who were traditionally certified. We do, however, find that teaching fellows were absent approximately one day more on average than other respondents (p-value = 0.03). No other traditional credentials were significant predictors of absences.

With regard to retention, we find negative effects for having a graduate degree (−0.06, p-value = 0.13) and being an education major (−.09, p-value = 0.03) on returning to teach in the DOE the following year, and positive effects for teaching fellows and TFA corps members (0.12 and 0.13, respectively, with p-values below 0.001). These results support the notion that teachers with more outside job opportunities are more likely to leave teaching in New York, but they may also reflect the particular nature of teaching fellows selection (in which commitment is a consideration) and the TFA program (for which there is an explicit two-year commitment).

Table 4 presents results on the nontraditional measures. All these measures have been normalized, so the coefficients can be interpreted as the estimated effect of moving 1 standard deviation in the distribution of the predictor. As in table 3, each row reflects impacts that are not conditional on any of the other predictors shown. That is, conditional on the school and student controls mentioned earlier, one can think of these as bivariate correlations between a single predictor and the outcome.

As hypothesized, the coefficients on these predictors are all positive, but they vary in size and statistical significance. Respondents' scores on the test of cognitive ability are not significant at conventional levels (p-value = 0.35), while math knowledge for teaching is more strongly related to math achievement, with a coefficient of 0.019, which is statistically significant at the 9 percent level. This gives support to the work by Hill, Rowan, and Ball (2005), who

21. This is a somewhat surprising result, and we can only speculate on its cause. It appears that this correlation exists only among the group of survey respondents who are white. However, among this group it is a fairly pervasive phenomenon; it exists within males, females, traditionally certified, and alternatively certified teachers, as well as within teachers whose college selectivity was above median and those whose college selectivity was below median.

Table 4. Nontraditional Predictors of Teacher and Student Outcomes

	Math Achievement	Subjective Evaluation	Teacher Absences	Returned to NYC	Returned to School | NYC
Cognitive ability (percentile, s.d. = 1)	0.011 (0.012) [0.348]	0.055 (0.062) [0.376]	−0.426 (0.232)* [0.067]	0.014 (0.017) [0.412]	0.020 (0.020) [0.322]
Math knowledge for teaching (percent correct, s.d. = 1)	0.019 (0.011)* [0.086]	0.020 (0.070) [0.773]	−0.401 (0.216)* [0.064]	0.006 (0.015) [0.685]	−0.014 (0.018) [0.439]
Conscientiousness (s.d. = 1)	0.015 (0.011) [0.156]	0.200 (0.066)** [0.003]	0.045 (0.168) [0.791]	−0.003 (0.015) [0.849]	0.006 (0.020) [0.780]
Extraversion (s.d. = 1)	0.001 (0.012) [0.943]	0.222 (0.065)** [0.001]	0.033 (0.190) [0.861]	−0.002 (0.016) [0.919]	0.025 (0.018) [0.169]
General efficacy (s.d. = 1)	0.015 (0.011) [0.191]	0.051 (0.061) [0.405]	0.031 (0.197) [0.877]	0.033 (0.017)* [0.058]	0.008 (0.017) [0.633]
Personal efficacy (s.d. = 1)	0.018 (0.010)* [0.075]	0.219 (0.065)** [0.001]	0.092 (0.204) [0.651]	0.017 (0.014) [0.245]	0.015 (0.016) [0.347]
Control for student/school characteristics and zip code FE	√	√	√	√	√
Observations	82,977	1,117	1,604	1,608	1,495

Notes: Each set of coefficients (separated by solid lines) represents different regressions. See text for a full listing of the student and school characteristics used as control variables. Standard errors (in parentheses) are clustered by school; *p*-values for each coefficient are shown in brackets. *significant at 10%; **significant at 5%; √ denotes controls included in the regression.

found this instrument to be a significant predictor of teacher effectiveness and a better predictor than other measures of teachers' math training.

The coefficient on conscientiousness (0.015) suggests that it has a moderate relation to student achievement, though the *p*-value is only .16, while the estimate for extraversion is quite close to zero (0.001). For general and personal efficacy, we also find positive, though imprecisely estimated, coefficients (0.015 with *p*-value of .19 and 0.018 with *p*-value of .08, respectively). Overall these results give mild support to the idea that teachers' personalities and attitudes are related to teacher effectiveness.[22]

22. We also test whether math achievement was higher among students assigned to teachers who placed greater emphasis on teaching skills related to test performance or who felt that the state standardized tests were good measures of students' knowledge and skills. As mentioned above, we collected these measures to try to address a concern that higher test score growth among students may simply reflect whether a teacher focuses on the test as an important outcome. However, the point estimates on both these variables are negative, with the coefficient on whether state tests are good measures of skills being statistically significant. It is not clear why students perform worse with

Interestingly, when we consider the relationship between these nontraditional measures and the subjective evaluations of teachers provided by mentors, we find noticeably different results. Subjective evaluations are higher for respondents with high levels of conscientiousness, extraversion, and personal efficacy, and the coefficients are quite large, ranging from 0.20 to 0.22, and highly statistically significant. In contrast, the evaluations bear little relation to the other nontraditional variables, though these coefficients are positive.

Given the somewhat contrasting results for math achievement and evaluations, it is important to point out that when subjective evaluations are used as a predictor of math achievement, we find that an increase of 1 standard deviation in the evaluation is associated with a 0.05 standard deviation increase in math test scores, which is a statistically and economically significant effect.[23] So while at least a portion of the variation in evaluations is based on observable differences in teachers' abilities to raise student achievement, another portion of the variance in evaluations is likely due to factors other than the ability to raise student test scores in math. We regard this as an important finding given the large literature on personality as a predictor of worker productivity, since most of these studies use subjective evaluations of employee performance by supervisors as the only outcome of interest.

With regard to absences, respondents with cognitive ability scores or math knowledge for teaching scores 1 standard deviation above average were absent 0.4 days less (p-value = .067). Respondents with general efficacy scores 1 standard deviation above average were more likely to return to the DOE (p-value = .058). As mentioned above, it is possible that responses to the efficacy instrument are influenced by the respondents' teaching experiences. At a minimum, this result then suggests that a teacher's willingness to stay in New York is correlated with feelings about self-efficacy.[24]

Overall, the results presented in tables 3 and 4 suggest that both traditional and nontraditional predictors may be associated with teachers' performance in their first year as measured by student achievement and teacher evaluations, absences, and turnover. However, there are a number of reasons to be cautious about these results. First, while most of the associations are in the expected direction, only a few are statistically significant. Given the large number of

teachers who believe the state tests are good measures of students' knowledge, but these estimates provide some support for the notion that teacher effectiveness as measured by value added on test scores is not simply an artifact of variation in the degree to which teachers focus on the skills measured by the tests.

23. Authors' calculations are available upon request. The use of these subjective evaluations by mentors as a means of identifying effective teachers after the recruitment stage is the subject of ongoing work by the authors (see Rockoff and Speroni 2010).

24. However, it is interesting that retention is related to general efficacy as opposed to personal, since the questions regarding personal efficacy are more focused on the teacher's own ability to succeed in the classroom.

coefficients being considered, any reasonable adjustment for testing multiple hypotheses would make these associations appear even less significant. Second, the fact that many of the coefficients are in the expected direction may reflect the predictors capturing similar underlying characteristics (so these estimates are not independent tests). Finally, the magnitudes of these effects, for math achievement in particular, are fairly modest relative to the differences that are known to exist across teachers. For example, Kane, Rockoff, and Staiger (2008) estimate a standard deviation of teacher effects on math achievement to be roughly 0.10 student-level standard deviations. Thus even the largest coefficient we estimate for math achievement (.019 on math knowledge for teaching) implies that we are predicting less than 4 percent of the teacher-level variation.

The Haberman PreScreener

The analysis above is largely exploratory, with the ultimate aim of identifying a variety of predictors that school officials might use to hire teachers who will be more effective in the classroom. As we noted earlier, there are several commercial teacher-screening instruments currently in use. In this section, we examine one of the most popular of such tools, the Haberman PreScreener (Haberman 1993, 1995). We first explore what characteristics and traits this instrument captures and then determine how well it predicts student and teacher outcomes.

Unlike the other nontraditional measures in our survey, the Haberman PreScreener is designed to evaluate a number of teacher characteristics simultaneously. Before we examine its relation to student and teacher outcomes, we use regression analysis to investigate how performance on this instrument is related to the demographic variables, traditional credentials, and nontraditional measures of teacher effectiveness included in tables 3 and 4. We focus on two dependent variables: the respondent's total score and whether the respondent placed in the "top group" using Haberman's scoring method. Haberman's approach to scoring candidates is somewhat complicated, taking into account not only the total score but also the presence of particularly low scores in any of the ten categories. Most important, any candidate who receives a score of "low" in one or more of the ten categories is automatically placed in the bottom quartile regardless of his or her total score.[25]

We present results that include each measure as a single predictor in separate regressions that also control for grade level taught and the school

25. For a more complete description of the Haberman scoring method, see an earlier version of this work (Rockoff et al. 2008). Description of the instrument and Haberman scoring method is based on personal communication with Martin Haberman and Delia Stafford in the fall of 2007 and subsequent conversations in the spring of 2008.

average characteristics from the CCD we used as control variables in tables 3 and 4. We use a probit regression for whether a respondent is in the top group and report marginal effects; results using ordinary least squares (OLS) are quite similar.

Performance on the Haberman PreScreener is significantly related to a number of these variables (table 5). Among the traditional credentials, performance on the Haberman is higher for respondents who passed the LAST on their first attempt and for those who have higher SAT verbal scores. Every nontraditional credential is positively related to performance on the Haberman PreScreener, and all save extraversion are statistically significant predictors of at least one of the two metrics.[26] Thus, as we expected, the questions on the Haberman PreScreener are designed to pick up on a number of the characteristics that prior research has put forth as predictors of teacher effectiveness.

We then use the same specification as for the other predictor variables to estimate the relationship between performance on the Haberman PreScreener and student achievement, subjective evaluations, absences, and retention (table 6). Again, we use two measures of performance: placing in the top group of candidates and total score. Generally we find stronger relationships when examining respondents' total scores than when examining whether a teacher would place in the top group. A 1 standard deviation increase in the score on the Haberman PreScreener is associated with a 0.022 standard deviation increase in math achievement that is marginally significant (p-value = 0.07) and a 0.16 standard deviation increase in subjective evaluation (p-value = 0.02). Increases in the score were also associated with a greater propensity to return to teaching the following year but also predicted a higher probability of transferring to another school within the DOE conditional on returning to teach. Both effects are only marginally significant (p-value = 0.2). While these results should be viewed with caution due to the timing of our survey, they lend some support to the notion that this instrument can identify characteristics that are correlated with teacher quality.

26. At first glance, it is somewhat puzzling that the results for being in the top group of candidates and the total score do not move in lockstep. However, it is important to recall that in order to be in the top group, candidates cannot have a low score on any of ten attributes. Because only a small subset of the fifty questions focuses on each attribute, it is quite possible to answer most questions correctly while still running afoul of this rule. In our sample, there are three attributes for which respondents were very likely to have a low score—Approach to Students (59 percent low), At-Risk Students (56 percent low), and Explains Teacher Success (50 percent low). Moreover, 69 percent of respondents scored low on at least one of these attributes, and there were no low scores on any attribute for the other 31 percent of our respondents. While the 69 percent of respondents with at least one low score had lower total scores than the other 31 percent of respondents, the difference—about four points—was only about 0.7 standard deviations in total score. Thus the distributions of total scores for these two groups overlap quite a bit.

Table 5. Predictors of Performance on the Haberman PreScreener

	In Top Group (Haberman Method) ***(marginal effects from probit)***	**Total Score** ***(s.d. = 1)*** ***(coefficient from OLS regression)***
Traditional credentials		
Has a graduate degree	0.078	0.013
	(0.057)	(0.134)
Passed LAST certification exam	0.167	0.352
on first attempt (1=yes)	(0.056)**	(0.239)
Teaching fellow	0.007	0.026
(relative to traditionally certified)	(0.061)	(0.137)
TFA corps member	−0.039	0.190
(relative to traditionally certified)	(0.074)	(0.168)
Math or science major (relative to majors	−0.094	−0.005
other than math, science, or education)	(0.068)	(0.174)
Education major (relative to majors	0.005	0.006
other than math, science, or education)	(0.061)	(0.138)
Self-reported SAT verbal score (s.d. = 1)	0.050	0.175
	(0.028)*	(0.062)**
Self-reported SAT math score (s.d. = 1)	−0.018	0.057
	(0.026)	(0.064)
Barron's rank of college (s.d. = 1)	0.029	0.069
	(0.032)	(0.071)
Nontraditional credentials		
Cognitive ability (percentile, s.d. = 1)	0.017	0.255
	(0.025)	(0.060)**
Math knowledge for teaching	0.049	0.198
(percent correct, s.d. = 1)	(0.024)**	(0.056)**
Conscientiousness (s.d. = 1)	0.052	0.026
	(0.024)**	(0.058)
Extraversion (s.d. = 1)	0.020	0.084
	(0.025)	(0.056)
General efficacy (s.d. = 1)	0.060	0.375
	(0.025)**	(0.053)**
Personal efficacy (s.d. = 1)	0.076	0.226
	(0.028)**	(0.066)**
Control for student/school characteristics	√	√

Notes: Each set of coefficients (separated by solid lines) represents different regressions where the outcome variable is regression on a single predictor or set of predictor variables. We use a probit regression to predict being in the top group according to Haberman's classification and report the mean marginal effect. We use least squares regressions to predict the total score and report coefficients. Robust standard errors are shown in parentheses.
*significant at 10%; **significant at 5%; √ denotes controls included in the regression.

Table 6. Haberman PreScreener Performance and Teacher and Student Outcomes

	Math Achievement	Subjective Evaluations	Teacher Absences	Returned to NYC	Returned to School | NYC
Haberman top group	0.047	0.296	0.787	0.003	−0.056
	(0.030)	(0.183)	(0.595)	(0.037)	(0.053)
	[0.124]	[0.107]	[0.187]	[0.932]	[0.284]
Haberman total score (s.d. = 1)	0.022	0.164	0.198	0.024	−0.026
	(0.012)*	(0.067)**	(0.232)	(0.019)	(0.020)
	[0.070]	[0.016]	[0.392]	[0.200]	[0.199]
Control for student/school characteristics and zip code FE	√	√	√	√	√
Observations	82,977	1,117	1,604	1,608	1,495

Notes: Each set of coefficients (separated by solid lines) represents different regressions. See text for a full listing of the student and school characteristics used as control variables. Standard errors (in parentheses) are clustered by school; *p*-values for each coefficient are shown in brackets. *significant at 10%; **significant at 5%; √ denotes controls included in the regression.

Factor Analysis and Predictions from Underlying Traits

The results presented above characterize the predictive power of various teacher characteristics taken individually. However, many of these elements are positively correlated and may serve as noisy measures of a small number of underlying traits. If so, combining several measures may yield a more reliable estimate of the underlying traits and thus provide more consistent predictive power for teacher and student outcomes. Therefore we estimate a factor model, which models all our measures as noisy estimates of a few underlying traits, and use the results to construct more reliable estimates of the underlying traits (the factors). We then use these estimated factors as predictors in a simplified analysis.

In the factor analysis, we include all the variables whose coefficients are shown in tables 3 and 4, as well as the Haberman total score.[27] The factor analysis results in two factors, which we call cognitive skills and noncognitive skills (see table 7). The six variables with the largest positive loadings on the first factor are all reasonable proxies for cognitive skills: being a TFA corps

27. The variables we include in the factor analysis are missing for some teachers, and traditional factor analysis fits the factor model to the correlation matrix constructed using only observations with complete data. In order to use all the available data, we instead estimated the factor analysis using the pairwise item correlation matrix. In total, we are able to measure these factors for 403 teachers. However, in results available upon request, we confirm that factors created using the listwise item correlation matrix and calculated only for the subset of teachers for whom all data are available produce comparable results. For ease of exposition, we apply a Promax rotation to the factor loadings, which produces factors that may be correlated with each other but maximize the extent to which each variable is associated with a single factor. To choose the number of factors, we use an eigenvalue cutoff of one, a commonly used standard in this methodology.

Table 7. Factor Analysis on Predictor Variables

Item	Factor 1: Cognitive Skills	Factor 2: Noncognitive Skills
Math or science major	0.0413	−0.2703
Teaching fellow	0.12	−0.4366
Teach for America	0.5732	0.2122
Passed LAST certification exam on 1st attempt (1 = yes)	0.2693	−0.0149
Barron's rank of college (s.d. = 1)	0.6043	−0.0845
Self-reported SAT math score (s.d. = 1)	0.6603	−0.15
Self-reported SAT verbal score (s.d. = 1)	0.6031	0.0182
Cognitive ability (percentile, s.d. = 1)	0.5527	−0.0793
Math knowledge for teaching (percent correct, s.d. = 1)	0.6441	−0.0091
Education major	−0.3422	0.234
Has a graduate degree	−0.183	0.1301
Extraversion (s.d. = 1)	0.0595	0.3655
Conscientiousness (s.d. = 1)	−0.1289	0.4398
Personal efficacy (s.d. = 1)	−0.1154	0.518
General efficacy (s.d. = 1)	0.4752	0.367
Haberman total score (s.d. = 1)	0.3029	0.3574

Note: Factor loadings calculated using the pairwise item correlation matrix and applying a Promax rotation.

member, attending a more selective college, SAT math score, SAT verbal score, cognitive ability as measured by the Raven's test, and math knowledge for teaching. The five variables with the largest positive loadings on the second factor are all reasonable proxies for other noncognitive skills important to teachers: extraversion, conscientiousness, personal efficacy, general efficacy, and the Haberman total score. Interestingly, being a teaching fellow (and, to a lesser extent, majoring in math or science) has considerable *negative* loadings on the noncognitive factor, while majoring in education has a considerable negative loading on the cognitive factor. It is worth noting that we obtain comparable results if, instead of the factors, we use a simple average of the measures with largest loadings on each factor.

In table 8, we use the predicted factors as predictive variables in regressions of student test scores and teacher-level outcomes, using the same specifications as with the single predictors but including both factors together. Both factors are positively and significantly associated with math achievement. Increasing

Table 8. Using Factors as Predictors of Teacher and Student Outcomes

	Math Achievement	Subjective Evaluations	Teacher Absences	Returned to NYC	Returned to School | NYC
Factor 1: Cognitive skills (s.d. = 1)	0.024 (0.010)**	0.021 (0.055)	−0.016 (0.050)	0.009 (0.004)**	0.002 (0.004)
Factor 2: Noncognitive skills (s.d. = 1)	0.025 (0.010)**	0.220 (0.052)**	−0.050 (0.055)	0.002 (0.004)	0.008 (0.005)*
F-test: All factors equal zero (*p*-value)	0.01	0.00	0.64	0.06	0.21
Observations	82,977	1,117	1,604	1,608	1,495
Control for student/school characteristics and zip code FE	√	√	√	√	√

Notes: All regressions include grade-level fixed effects, school zip code fixed effects, and student-, class-, and school-level observable characteristics (see text for a complete list). Standard errors (in parentheses) are clustered at the school level.
*significant at 10%; *significant at 5%; √ denotes controls included in the regression.

cognitive or noncognitive skills by one point is associated with increases in student achievement of 0.024 and 0.025 standard deviations, respectively. Interestingly, only noncognitive skills have a significant positive relationship with subjective evaluations. Cognitive skills have a significant positive association with retention within the DOE, while noncognitive skills have a significant positive association with retention within a school, conditional on returning to the DOE.[28]

The effects of cognitive and noncognitive skills on student achievement are modest but still economically important. Moreover, our ability to measure these two sets of skills is greatly improved by the use of the nontraditional measures gathered in our survey. To illustrate both of these points, we take the estimates from column 1 of table 8 and assign each teacher respondent the predicted impact on student achievement associated with these two factors. We also estimate the cognitive and noncognitive factors using only the traditional credentials (i.e., we act as if the nontraditional measures were unavailable for our survey respondents), repeat our regression analysis, and again predict impacts for respondents. We then plot the distributions for these two sets of estimates in figure 1. For additional comparisons, we also plot a simulated

28. In results not reported here, we recreated the factors and reestimated the regressions shown in table 8, omitting the self-efficacy measures because they may suffer from reverse causality, as explained above. The results are virtually identical. We also examine the heterogeneity of the relationships between the factors and student achievement by school-level (elementary versus middle) and school poverty (below versus above median free lunch receipt) characteristics. We find some evidence that cognitive skills were more important in low-poverty schools (*p*-value = 0.18) and noncognitive skills were more important in elementary schools (*p*-value = 0.17).

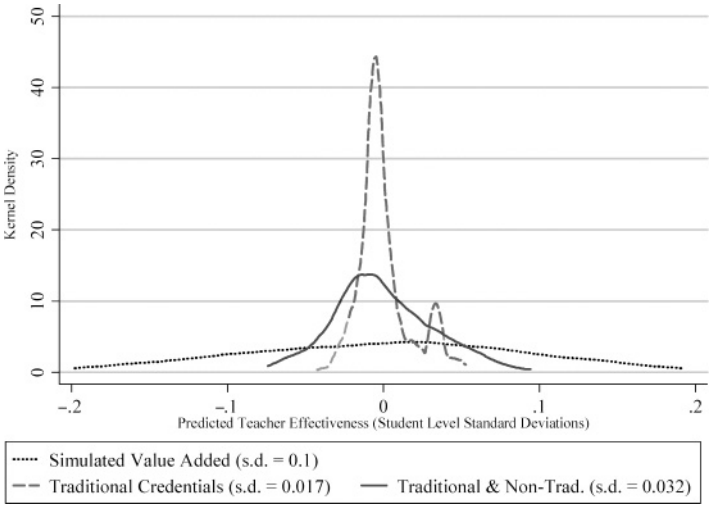

Figure 1. Recruitment Information and the Distribution of Predicted Value Added. Notes: Kernel density plots are shown of predicted value added from two regressions of student test scores on a set of teacher characteristics and other controls. "Simulated value added" is the kernel density plot of a randomly drawn normally distributed random variable with mean 0 and standard deviation 0.10. See text for a list of regression covariates.

distribution of teacher effectiveness, which is simply a normal distribution with a standard deviation of 0.10. This approximates the variation in value added among new teachers estimated by Kane, Rockoff, and Staiger (2008) for New York City teachers and serves as a simple benchmark against which to measure the variation in predicted teacher effectiveness using the two factors.

Examining these plots, we see a clear increase in the variation of predicted teacher effectiveness as we use the information from nontraditional credentials (figure 1). The standard deviation of predicted teacher effectiveness using only the traditional credentials to generate our factor estimates is 0.017, and adding the nontraditional credentials raises the standard deviation to 0.032.[29] This suggests that districts may be able to gain some traction in selecting more effective teachers by using broader sets of information during recruitment. However, the variation of predicted value added with an expanded set of data on new teachers has only about 10 percent of the variance of the expected distribution of teacher effectiveness. This underscores the difficult—perhaps

29. The bump in the distribution of predicted effectiveness based on traditional characteristics, shown in figure 1, is driven primarily by higher predicted effectiveness of TFA corps members. Also, note that we might have plotted predictions of teacher effectiveness using regressions that included all the individual credentials as covariates. However, a large number of variables capturing information on teachers would be able to explain some variation in student achievement even if these variables were completely invalid predictors of teacher effectiveness. Indeed, using Monte Carlo simulations, we find that random assignment of a large number of characteristics (e.g., 10–15) generates substantial variance in predicted effectiveness, on the order of 0.06 to 0.08 standard deviations.

impossible—task of identifying systematically the most highly effective or ineffective teachers without any data on actual performance in the classroom.

5. CONCLUSION

We use a survey of new teachers in New York City to investigate whether one can predict economically significant variation in teacher effectiveness using a broadened set of information on new recruits. The evidence we present suggests that this is the case, and it shows in particular that predictive power is gained by using measures of teacher effectiveness suggested by earlier research but rarely, if ever, collected and used by school districts.

Our findings are in a spirit similar to a recent article by Boyd et al. (2008), which makes the argument that recruiting teachers with a number of attractive credentials while avoiding teachers whose credentials are unattractive has the potential power to improve the effectiveness of their teacher workforce. Importantly, their results rely not on any single variable (e.g., teacher certification pathway) but instead on a broad set of credentials, all of which are fairly traditional indicators of teacher quality, but some (e.g., SAT scores) are not currently collected by many school districts, including New York City. Our results go further and suggest collecting a set of measures that would not appear on a teacher's curriculum vitae.

While our findings provide motivation for schools to expand the set of criteria used in recruitment, there are a number of reasons why the results should be interpreted with caution. First, our survey was completed well after the start of the school year. Thus teachers' experiences during the school year may have affected some of their responses. For most survey items, the problem of reverse causality is highly unlikely (e.g., reported SAT scores or cognitive ability), but for others it may be potentially important (e.g., feelings on personal efficacy). Second, the only way to truly validate our findings is to gather a similar set of information on a new sample of teachers and test whether our results here are also found for this new sample. Thus more work is necessary in this line of research.

The authors would like first to thank Jon Fullerton, who helped us greatly in the design and implementation of the survey used in this analysis. We also thank a number of individuals who made the survey possible, including Betsy Arons, Vicki Bernstein, Nate Brown, Doug Jaffe, Leigh McGuigan, Amy McIntosh, Joe Meglino, and Ranjeet Singh of the New York City Department of Education, Delia Stafford and Martin Haberman of the Haberman Foundation, and Heather Hill of the Harvard Graduate School of Education. Ellen Viruleg, Stephanie Rennane, and Robert Lindsley provided outstanding research assistance. We are grateful to the Spencer Foundation and the Carnegie Corporation for generous financial support.

REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25(1): 95–135.

Armor, David, Patricia Conry-Oseguera, Millicent Cox, Nicelma King, Lorraine McDonnell, Anthony Pascal, Edward Pauly, and Gail Zellman. 1976. *Analysis of the school preferred reading program in selected Los Angeles minority schools*. Santa Monica, CA: RAND Corporation.

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.

Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2006. How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy* 1(2): 176–216.

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah E. Rockoff, and James Wyckoff. 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.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources* 41(4): 778–820.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2007. How and why do teacher credentials matter for student achievement? NBER Working Paper No. 12828.

Decker, Paul T., Daniel P. Mayer, and Steven Glazerman. 2004. The effects of Teach for America on students: Findings from a national valuation. Mathematica Policy Research Report No. 8792–750.

Dembo, Myron H., and Sherri Gibson. 1985. Teachers' sense of efficacy: An important factor in school improvement. *Elementary School Journal* 86(2): 173–84.

Getzels, Jacob W., and Phillip W. Jackson. 1963. The teacher's personality and characteristics. In *Handbook of research on teaching*, edited by Nathaniel L. Gage, pp. 506–82. Chicago: Rand McNally.

Gibson, Sherri, and Myron H. Dembo. 1984. Teacher efficacy: A construct validation. *Journal of Educational Psychology* 76(4): 569–82.

Goldhaber, Dan D. 2007. Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources* 42(4): 765–94.

Goldhaber, Dan D., and Dominic J. Brewer. 1997. Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources* 32(3): 505–23.

Goodstein, Leonard D., and Richard I. Lanyon. 1999. Applications of personality assessment to the workplace: A review. *Journal of Business and Psychology* 13(3): 291–322.

Gordon, Robert, Thomas Kane, and Douglas O. Staiger. 2006. The Hamilton Project: Identifying effective teachers using performance on the job. Washington, DC: Brookings Institution.

Greenwald, Rob, Larry V. Hedges, and Richard D. Laine. 1996. The effect of school resources on student achievement. *Review of Educational Research* 66(3): 361–96.

Haberman, Martin. 1993. Predicting the success of urban teachers (the Milwaukee trials). *Action in Teacher Education* 15(3): 1–5.

Haberman, Martin. 1995. Selecting "star" teachers for children and youth in urban poverty. *Phi Delta Kappan* 76(10): 777–81.

Hanushek, Eric A. 1986. The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature* 24(3): 1141–77.

Hanushek, Eric. A. 1997. Assessing the effects of school resources on student performance: An update. *Educational Evaluation and Policy Analysis* 19(2): 141–64.

Harris, Douglas N., and Tim R. Sass. 2006. The effects of teacher training on teacher value added. Working Paper No. WP_2006_03_01, Florida State University.

Harris, Douglas N., and Tim R. Sass. 2008. What makes for a good teacher and who can tell? Unpublished paper, Florida State University.

Hill, Heather C., Brian Rowan, and Deborah L. Ball. 2005. Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal* 42(2): 371–406.

Hoy, Wayne K., and Anita E. Woolfolk. 1993. Teachers' sense of efficacy and the organizational health of schools. *Elementary School Journal* 93: 356–72.

Jacob, Brian A. 2007. The challenges of staffing urban schools with effective teachers. *Future of Children* 17(1): 129–53.

Jacob, Brian A., and Lars J. Lefgren. 2008. Principals as agents: Subjective performance measurement in education. *Journal of Labor Economics* 26(1): 101–36.

John, Oliver P., Eileen M. Donahue, and Robert L. Kentle. 1991. The "big five" inventory—versions 4a and 54. Berkeley: University of California, Institute of Personality and Social Research.

Kane, Thomas, 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(6): 615–31.

Murnane, Richard J. 1975. *The impact of school resources on the learning of inner city children*. Cambridge, MA: Ballinger Publishing Co.

Raudenbush, Stephen W., Brian Rowan, and Yuk Fai Cheong. 1992. Contextual effects on the self-perceived efficacy of high school teachers. *Sociology of Education* 65(2): 150–67.

Raven, John C. 2000. The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology* 41(1): 1–48.

Raven, John C., John H. Court, and John Raven. 2000. *Manual for Raven's progressive matrices and vocabulary scales. Section 3, standard progressive matrices*. San Antonio, TX: Harcourt Assessment.

Raven, John, and W. A. Summers. 1986. A compendium of North American normative and validity studies. In *Manual for Raven's progressive matrices and vocabulary tests*, edited by John C. Raven, John H. Court, and John Raven, Research Supplement No. 3. San Antonio, TX: Psychological Corporation.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools, and academic achievement. *Econometrica* 73(2): 417–58.

Rockoff, Jonah E. 2004. The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review* 94(2): 247–52.

Rockoff, Jonah E. 2008. Does mentoring reduce turnover and improve skills of new employees? Evidence from teachers in New York City. NBER Working Paper No. 13868.

Rockoff, Jonah E., Brian A. Jacob, Thomas J. Kane, and Douglas O. Staiger. 2008. Can you recognize an effective teacher when you recruit one? NBER Working Paper No. 14485.

Rockoff, Jonah E., and Benjamin B. Lockwood. 2010. Stuck in the middle: Educational impacts of grade configuration. *Journal of Public Economics* 94(11–12): 1051–61.

Rockoff, Jonah E., and Cecilia Speroni. 2010. Subjective and objective evaluations of teacher effectiveness. *American Economic Review* 100(2): 261–66.

Shulman, Lee S. 1986. Those who understand: Knowledge growth in teaching. *Educational Researcher* 15(2): 4–14.

Shulman, Lee S. 1987. Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review* 57(1): 1–22.

Wilson, S. M., Lee S. Shulman, and A. Richert. 1987. 150 different ways of knowing: Representations of knowledge in teaching. In *Exploring teachers' thinking*, edited by James Calderhead, pp. 104–24. London: Cassell Educational.

Woolfolk, Anita E., and Wayne K. Hoy. 1990. Prospective teachers' sense of efficacy and beliefs about control. *Journal of Educational Psychology* 82(1): 81–91.